Conserved basic residues in the C-type lectin and short complement repeat domains of the G3 region of proteoglycans

Nigel C. BRISSETT and Stephen J. PERKINS¹

Department of Biochemistry and Molecular Biology, Royal Free Hospital School of Medicine, Rowland Hill Street, London NW3 2PF, U.K.

Aggrecan is the major proteoglycan of the extracellular matrix in cartilage. It contains two N-terminal globular regions, G1 and G2, and one C-terminal globular region, G3. G3 is implicated in the intracellular processing of aggrecan and contains a C-type lectin carbohydrate recognition domain (CRD), frequent occurrences of a C-terminal short complement repeat (SCR) domain, and occasionally an N-terminal epidermal growth factor domain. The CRD and SCR domains in 13 G3 sequences were each subjected to structural analysis. Alignment of 131 sequences from all seven groups in the CRD superfamily defined a consensus length of 136 residues, in which 32% of residues were conserved. Although the G3 CRD sequences agreed with this consensus, they also contained five fully conserved basic residues that are atypical of the CRD superfamily. Homology modelling showed that four of these residues are located on a surface region on the

CRD that is separate from the Ca²⁺-binding residues involved in carbohydrate interactions. One conserved basic residue is identical in position with that of a conserved basic residue that mediates hyaluronate binding in the structurally related proteoglycan tandem repeat (PTR) domain in G1 and in link protein. The alignment of 13 G3 SCR sequences with 101 sequences in the SCR superfamily showed good agreement with conserved residues in the SCR superfamily. There are also five conserved basic residues in the G3 SCR that are atypical of the SCR superfamily, and homology modelling showed that all five were located on one surface of the SCR. It is concluded that both the CRD and SCR domains in G3 possess basic residues that are atypical of their superfamilies and might be related to function, and that the G3 CRD domain shows an evolutionary relationship to the PTR domain in G1.

INTRODUCTION

The aggregating group of proteoglycans consists of many long anionic polysaccharide chains (glycosaminoglycans) covalently attached to an extended protein core. Proteoglycans form aggregates with hyaluronate and, together with collagen, maintain the tensile and elastic properties of cartilage [1,2]. Aggrecan, versican, neurocan and brevican are four members of the proteoglycan family. Aggrecan contains two N-terminal globular regions, G1 and G2, with immunoglobulin (Ig) fold and proteoglycan tandem repeat (PTR) domains (Figure 1). Link protein contains the same three-domain structure as G1, and with G1 forms a very stable ternary complex with hyaluronate. G1 and G2 are followed by a large extended intermediate region of 1487 residues containing keratan sulphate and chondroitin sulphate attachment sites. The C-terminal G3 region is present when aggrecan is secreted but is lost as cartilage ages [3–5]. G3 has been implicated in the intracellular synthesis and maturation of aggrecan, as its absence causes nanomelia, a lethal genetic mutation in chicken [6–8]. G3 contains one of two alternatively spliced epidermal growth factor domains (EGF1 and EGF2), followed by a C-type lectin carbohydrate recognition domain (CRD) and a short complement/consensus repeat (SCR) domain. The CRD is always present, whereas the occurrence of SCR domains in G3 is variable [9,10] and those for the EGF1 and EGF2 domains are 25–28% and 5–8% [11].

The CRD and SCR domains belong to large superfamilies with known protein folds. The G3 CRD domain corresponds to group I of the seven groups I to VII in the CRD superfamily of

length 115–130 residues that have been analysed by Drickamer [12–14]. Crystal structures are known for mannose-binding protein A and C, E-selectin, lithostathine and tetranectin in groups III, IV and VII [15–21]. The CRD contains a characteristic motif of two α -helices between which are three antiparallel β -strands (β), this motif being attached to three further β -strands in an upper region with two Ca²⁺-binding sites [22,23]. A related $2\alpha/3\beta$ motif and a truncated upper region is found in the PTR domain [23–25]. The G3 SCR domain is part of the SCR superfamily, the most abundant domain type in the complement proteins, with length 60–62 residues. NMR structures are known for two SCR domains in human complement factor H [26,27], these being formed from three pairs of antiparallel β -strands [28,29].

Structures for the CRD and SCR domains in G3 are required for an understanding of possible functions for G3 in cartilage and for elucidating any evolutionary relationship between the CRD and PTR domains in aggrecan (Figure 1). In the absence of crystal structures, the availability of large sequence superfamilies and known protein folds for the CRD and SCR domains constitutes a basis for homology modelling of these structures in G3. Homology modelling has previously been applied to CRD domains from groups II and VII, namely CD23/Fc ϵ RII, CD69 and type II antifreeze protein [30–32], and alignments of up to 50 CRD sequences have been reported [13,14]. Here we present a multiple sequence alignment based on 131 CRD sequences. In combination with another alignment based on 101 SCR sequences, together with consensus secondary structure predictions and averaged protein fold recognition analyses, we present homology

Abbreviations used: CRD, carbohydrate recognition domain of the C-type lectin superfamily; EGF, epidermal growth factor; G1, G2 and G3, three globular regions of aggrecan; PTR, proteoglycan tandem repeat; SCR, short complement/consensus repeat.

¹ To whom correspondence should be addressed.

Figure 1 Domain structure in the G1, G2 and G3 regions of human aggrecan

Each domain is denoted by a circle. The residue numbering below the domains indicates the size of each domain and the length of the polypeptide chain between them. A large region with negatively charged keratan sulphate (KS) and chondroitin sulphate (CS) glycosaminoglycans occurs between G2 and G3. The location of putative N-glycosylation sites is indicated above the domains.

models for the G3 CRD and SCR domains in human aggrecan. Both homology models revealed unexpected sets of conserved basic residues within the G3 region, both of which are atypical of the CRD and SCR superfamilies and might be of functional significance. There is currently much interest in the ability to predict an unknown protein fold starting from a sequence alignment [33], so we also appraise the applicability of these methods to both the CRD and the PTR domains.

METHODS

Sequence retrieval and alignment, consensus structure predictions and protein fold recognition

The procedures and programs used in this study were as described in [23,34]. A total of 131 CRD sequences were extracted from ENTREZ (National Center for Biotechnology Information). These totalled 13 from group I (proteoglycans), 19 from group II (type II receptors), 15 from group III (collectins), 13 from group IV (selectins), 10 from group V (type II lymphocyte antigens), 24 from group VI (mannose receptors), and 37 from group VII (miscellaneous), with SWISSPROT or PIR accession names or numbers summarized in Figures 2 and 3. The CRD domain boundaries were taken from ENTREZ when specified, but were otherwise arbitrary. The sequence alignment was initiated with the multiple alignment program MULTAL with a range of fixed and variable gap penalties. A final refinement of the alignment maximized the occurrence of conserved or chemically similar residues, and minimized the occurrence of sequence gaps to facilitate the calculation of the consensus predicted structure. Residues were defined into the subgroups tiny (G, A, S), aliphatic (I, L, V, M), aromatic (H, F, Y, W), positively charged (R, H, K), negatively charged (D, E), hydroxy (S, T) and amide (N, Q). On the basis of the alignment, five averaged three-state or four-state secondary structure predictions were based on the classical GOR I/GOR III and Chou–Fasman statistical methods [28], the environment-dependent amino acid substitution probability method SAPIENS, and the neural networking PHD method. The mean hydropathy was predicted by using a consensus hydrophobicity score [28]. Mean solvent accessibilities were predicted by the SAPIENS and PHD approaches. A similar procedure was performed for the 13 G3 SCR sequences. The 131 CRD sequences were used in the optimal fitting of each sequence to a library of 254 protein folds by using the pairwise interaction

energies in THREADER. The threadings were represented as Z-scores [$= (\text{energy} - \text{mean}) / \text{S.D.}$]. A good structural match requires a Z-score of less than -2.7 . The Z-scores were sorted for input into SUM_THREADER to calculate the average Z-score and position of each of the 254 folds in the 131 threadings [34].

Protein molecular modelling

Protein structures were visualized with INSIGHT II 95.0 (Biosym/MSI, San Diego, CA, U.S.A.) and SETOR on Silicon Graphics INDY Workstations. The G3 CRD of human aggrecan was modelled by the rigid body fragment assembly method in HOMOLGY (Biosym/MSI) by using the crystal structure of mannose-binding protein [15]. The Brookhaven database code is 2msb; those for other structures are denoted by four-letter codes in parentheses. The structurally conserved core region corresponded to 28 α -helix and β -sheet residues and 75 loop residues. Two surface loops between Asn-102 and Glu-106 (five residues) and Met-112 and Gly-117 (five residues) in the G3 CRD were modelled by using the Brookhaven loop database of INSIGHT II. Two C-terminal residues (Gly-135 and Thr-136) were added. The N-terminal extension of the G3 CRD model starting from 2msb was modelled by adding the first 14 residues (Asp-7–Gly-20) in an extended β -strand conformation, then introducing two β -turns at Glu-12 and Tyr-18 to join Cys-11 and Cys-22. Energy refinements with DISCOVER (Biosym/MSI) were performed first at the loop–splice junctions, then on the loops, then on the mutated core residues. The 103 core residues were tethered to their original positions with greater energy terms than the reconstructed loops with 26 residues. Refinements were based on the consistent valence force field, and iterations were made by using combinations of the steepest descent and conjugate algorithms. The connectivity of the model was improved and the proportion of bad contacts or stereochemistry was decreased, as confirmed by the use of PROCHECK. The G3 SCR of human aggrecan was modelled similarly by using the NMR structure of SCR-15 and SCR-16 in human factor H [27] (1hfh). The conserved protein core was taken to be 15 β -sheet and 33 loop residues, the surface loop residues between Lys-21 and Glu-25 (three residues), Cys-48 and His-53 (four residues) and Trp-54 and Thr-60 (five residues) were modelled from the loop database, and the end residues (Gly-1 and Thr-62) were added.

The observed secondary structure in crystal or NMR structures was assigned by using the DSSP program (nomenclature: β -strands, E and B; α -helices, I, H or G; loop regions, s, t and .). Side-chain solvent accessibilities were calculated by the Lee and Richards method in the COMPARE program on a scale from 0 to 9 for each residue, where 0 corresponds to 0–10% solvent exposure, 1 corresponds to 11–20% solvent exposure, and so on. By using DELPHI (Biosym/MSI), the total electrostatic surface was calculated with a full coulombic boundary condition, with the interior and exterior dielectric constants set as 2 and 80 respectively and an ionic strength corresponding to 0.145 M salt. The Connolly algorithm with a solvent probe radius of 0.14 nm was used to create the protein surface, which was coloured red for acidic potentials less than $-2kT$ and blue for basic potentials greater than $2kT$.

RESULTS AND DISCUSSION

Sequence alignment and residue conservation in the CRD superfamily

Crystal structures are known for mannose-binding protein, E-selectin and lithostathine in the CRD superfamily [15–21]. To correlate these with the G3 CRD structure, an alignment of 131 CRD sequences is reported in both Figures 2 and 3. The alignment showed that the consensus sequence length of the CRD is 136 residues. Previous alignments were based on up to 50 sequences and resulted in the definition of seven CRD groups [13,14] with either three or two conserved disulphide bridges, which are here termed the long and short CRD subsets respectively. As previously, the long CRDs are represented by 95 sequences in groups I, II, V, VIb and VII and contain two extra N-terminal cysteine residues Cys-11 and Cys-22 that are not found in the short CRDs of groups III, IV and VIa. The greater number of CRD sequences results in a better definition of residue conservation than previously, and this is low. Only 26 of 136 residue positions were conserved or conservatively replaced in over 70% of the 131 sequences (Figure 2). The 10 residues conserved in over 90% of the sequences are four cysteines (Cys-39, Cys-109, Cys-124 and Cys-132), five hydrophobic residues (Tyr/Phe-23, Trp-32, Leu-46, Trp-70, Ile-71) and Gly-72. The CRD crystal structures are constructed from a $2\alpha/3\beta$ motif at the base of the structure, which is formed from the α -helices A1 and A2 and the β -strands B1, B2 and B7, and this motif makes contact with B3 in the neighbouring β -sheet (Figure 3) [23]. With the exception of Cys-109 and Cys-124, eight of these ten best-conserved residues are found within this $2\alpha/3\beta$ motif and B3, and define the conserved structural core of the CRD. Further inspection of both the short and long CRD subsets shows that the remaining 16 residues that are conserved in 70% of sequences (Figures 2 and 3) are distributed throughout the structure.

When the G3 CRD sequences were compared with the rest of the CRD superfamily, a total of 24 of these 26 well-conserved residues were also found as expected within the 13 G3 CRD sequences (Figure 3). Within the G3 CRD sequences, 73/129 residues are conserved to better than 90%. The comparison of the conserved G3 CRD residues with the CRD consensus in Figure 3 shows that five fully conserved basic residues occur at Arg/Lys-24, Arg-29, Arg-40, Arg/Lys-76 and Lys-133. These have no counterpart in the 118 other CRD sequences of Figure 2. In these 118 sequences, only 3 of these 5 residues occur in 8 sequences, 2 in 27 sequences, 1 in 52 sequences, and 0 in 31 sequences, which is as expected on the basis of random events. The CRD fold is structurally related to the PTR fold [23]. Comparison of Figure 3 with an alignment of 59 PTR sequences [23] shows that Arg-29 in the G3 CRD sequences is structurally

analogous to Lys-11 at the C-terminus of β -strand B1 in the related $2\alpha/3\beta$ motif that occurs in the PTR fold. The conserved basic residue at Arg-9/Lys-11 has been implicated in hyaluronate binding to the PTR of the cell surface proteoglycan CD44 [35]. This similar occurrence of a conserved basic residue at the same position in both folds suggests that there is an evolutionary relationship between the PTR and the G3 CRD.

In mannose-binding protein, CRD carbohydrate binding occurs through hydrogen bond formation with a Ca^{2+} ion co-ordinated at site 2 at Glu-99–Asn-101–Glu-107 (or Asn-102 in E-selectin)–Asn-120–Asp-121 (Figure 3). Mannose specificity is determined primarily by Glu-99–Asn-101; galactose specificity is determined by their replacement to Gln-99–Asp-101. A Ca^{2+} site 1 is found at Asp-75–Glu-79–Asp-102–Asp-108 in mannose-binding protein that does not exist in E-selectin. These nine residues are usually conserved in the G3 CRD sequences. Because they are located in an extensive loop region between residues 75 and 127 (Figure 3), it is not surprising that the CRD structure is sensitive to residue replacements in this region [36]. The residues DHGSG and NRQKD that follow immediately after Glu-99–Asn-101 in mannose-binding protein and E-selectin respectively become NFFAAG in the G3 CRD (Figure 3). Similarly IS and ER immediately after Asn-120–Asp-121 become VP in the G3 CRD. Both these replacements are more hydrophobic and might affect carbohydrate binding.

Observed and predicted protein structures in the CRD superfamily

The observed secondary structure in CRD crystal structures was compared with the outcome of five consensus secondary-structure predictions based on the CRD sequence alignment [22,23]. The DSSP analyses of five crystal co-ordinate sets (Brookhaven codes 1msb, 2msb, 1esl) showed that two α -helices A1 and A2 and seven β -strands B1 to B7 were consistently observed. A short β -strand B2 that is part of the $2\alpha/3\beta$ motif was observed that had not been described in the original crystallographic reports. Comparison with the consensus prediction [22,23] showed that two α -helices and six β -strands were correctly predicted. The exception was B6, which contains part of Ca^{2+} -binding site 2 in mannose-binding protein and E-selectin. This is readily attributed to the conserved residues Asn-120 and Asp-121 on B6, as these residues do not favour β -strand conformations in proteins ([37], Table 6.5), yet are required for Ca^{2+} -binding. A total of 98/119 matches occurred between the observed and predicted structures, i.e. a prediction accuracy of 82%. In the present study, a consensus prediction based on the 13 G3 CRD sequences showed a decreased accuracy of 66% (78/116). Although this lower accuracy is attributable to the fewer sequences used, the prediction showed that the main features of the G3 CRD were all present as expected (A1, A2, B2, B3, B5 and B7). Comparison of the observed and predicted CRD residue solvent accessibilities showed that the prediction accuracy was 77% (92/119 matches) [23]. Here, although the G3 CRD sequences gave a lower overall accuracy of 62% (72/116 matches), good matches were seen for the main secondary structure elements as expected (A1, A2, B1, B2, B3, B5 and B7).

Protein fold recognition analyses were performed for the CRD superfamily by using THREADER. Each of the 131 CRD sequences was threaded through a set of 254 known protein folds, which included mannose-binding protein (2msb), and scored for topological compatibility with each fold by using a set of inter-residue energy potentials. The averaged result for all 131 CRD sequences showed that 2msb appeared first out of 254 folds with a mean Z-score of -2.17 ± 1.31 . For the 36 short CRD sequences, a notably high averaged Z-score of -2.92 ± 0.60 was

[illegible]

Figure 2 For legend see facing page

obtained with the 2msb fold, which also scored best. For the 95 long CRD sequences, a lower averaged Z-score of -1.89 ± 1.40 was obtained with the 2msb fold, and this came second after the all- β protein interleukin 1 β . For the 13 G3 CRD sequences, a similar averaged Z-score of -1.87 ± 0.60 was obtained with the

2msb fold, and this was in fourth place out of 254. As the secondary structures of the three higher-scoring folds in this case were clearly inconsistent with the predicted secondary structure, the outcome of **THREADER** favoured the 2msb fold as the closest structure to that of the G3 CRD fold. At present the

CONSENSUS Group I CRD SEQUENCES														10	20	30	40	50	60	70	80	90	100	110	120	130																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
PGCA_HUMAN (Aggrecan: human)														DQEVCEEGN	KYQGH	CYRHP	DPRET	WDAER	CRREQ	SHSS	IVTPE	EQEFVN	NAQDY	QWIGL	NDRTIE	EGDFR	WSDGHP	MQ	FENWR	PNQDN	FFA	GEDC	VVM	IWHE	KGEW	NDVPC	NHLP	FTCK	KGT																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																
PGCA_RAT (Aggrecan: rat)														DQEQCEEG	TKFG	HCYR	HPDRE	TWDAE	RCREQ	SHSS	IVTPE	EQEFVN	NAQDY	QWIGL	NDRTIE	EGDFR	WSDGHS	LQ	FEKWR	PNQDN	FFA	GEDC	VVM	IWHE	RGEW	NDVPC	NHLP	FTCK	KGT																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																
PGCA_BOVIN (Aggrecan: bovine)														DQKLCEEG	TKFG	HCYR	HPDRE	TWDAE	RCREQ	SHSS	IVTPE	EQEFVN	NAQDY	QWIGL	NDRTIE	EGDFR	WSDGHS	LQ	FENWR	PNQDN	FFA	GEDC	VVM	IWHE	KGEW	NDVPC	NHLP	FTCK	KGT																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																
PGCA_CHICK (Aggrecan: chick)														DLANCEEG	IKFG	HCYR	HPDRE	TWDAE	RCREQ	SHSS	IVTPE	EQEFVN	NAQDY	QWIGL	NDRTIE	EGDFR	WSDGHS	LQ	FEKWR	PNQDN	FFA	GEDC	VVM	IWHE	RGEW	NDVPC	NHLP	FTCK	KGT																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																
A55182 (Aggrecan: mouse)														DQEQCEEG	TKFG	HCYR	HPDRE	TWDAE	RCREQ	SHSS	IVTPE	EQEFVN	NAQDY	QWIGL	NDRTIE	EGDFR	WSDGHS	LQ	FEKWR	PNQDN	FFA	GEDC	VVM	IWHE	KGEW	NDVPC	NHLP	FTCK	KGT																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																
A54423 (Brevican: bovine)														GLHFCSP	GMWAF	QGAC	YKHF	SARRS	WEEA	ENKCR	MYGA	HLAS	ISTP	EEQDF	INNR	YREY	QWIGL	NDRTIE	EGFL	WSDG	PVLL	YENWN	PGQDS	YFL	SGENC	VVM	WHD	QGQW	SDVPC	NHLY	STCK	MGL																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
A54126 (Brevican: rat)														GLHFCSP	GMWAF	QGAC	YKHF	SARRS	WEEA	ENKCR	MYGA	HLAS	ISTP	EEQDF	INNR	YREY	QWIGL	NDRTIE	EGFL	WSDG	PVLL	YENWN	PGQDS	YFL	SGENC	VVM	WHD	QGQW	SDVPC	NHLY	STCK	MGL																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
B86890 (Brevican: mouse)														GLHFCSP	GMWAF	QGAC	YKHF	SARRS	WEEA	ENKCR	MYGA	HLAS	ISTP	EEQDF	INNR	YREY	QWIGL	NDRTIE	EGFL	WSDG	PVLL	YENWN	PGQDS	YFL	SGENC	VVM	WHD	QGQW	SDVPC	NHLY	STCK	MGL																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
S28764 (Neurocan: rat)														DTEGCDRG	WGH	QCYR	YFAH	RWDAE	RCRL	QGAHL	TS	VSHP	PEEK	IF	NSF	GEN	SWT	GLND	TR	VER	FQ	MT	DT	GL	YEN	WR	KQ	PN	DF	AG	ED	CV	TI	WHE	SR	GN	WD	VP	CP	NY	LP	VC	CK	GT																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																	
S52781 (Neurocan: mouse)														DTEGCDRG	WGH	QCYR	YFAH	RWDAE	RCRL	QGAHL	TS	VSHP	PEEK	IF	NSF	GEN	SWT	GLND	TR	VER	FQ	MT	DT	GL	YEN	WR	KQ	PN	DF	AG	ED	CV	TI	WHE	SR	GN	WD	VP	CP	NY	LP	VC	CK	GT																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																	
PGCS_HUMAN (Versican: human)														DTETCDY	GMWAF	QGAC	YKHF	SARRS	WEEA	ENKCR	MYGA	HLAS	ISTP	EEQDF	INNR	YREY	QWIGL	NDRTIE	EGDFR	WSDG	PVLL	YENWN	PGQDS	YFL	SGENC	VVM	WHD	QGQW	SDVPC	NHLY	STCK	MGL																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
A47171 (Versican: chick)														DTETCDY	GMWAF	QGAC	YKHF	SARRS	WEEA	ENKCR	MYGA	HLAS	ISTP	EEQDF	INNR	YREY	QWIGL	NDRTIE	EGDFR	WSDG	PVLL	YENWN	PGQDS	YFL	SGENC	VVM	WHD	QGQW	SDVPC	NHLY	STCK	MGL																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
B62461 (Versican: mouse)														DTETCDY	GMWAF	QGAC	YKHF	SARRS	WEEA	ENKCR	MYGA	HLAS	ISTP	EEQDF	INNR	YREY	QWIGL	NDRTIE	EGDFR	WSDG	PVLL	YENWN	PGQDS	YFL	SGENC	VVM	WHD	QGQW	SDVPC	NHLY	STCK	MGL																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
Most frequent residue (>90% conservation)														C	GN	FQG	CYRHF	R	W	AE	CR	AHLTSI	EE	FVN	D	WIGLND	R	E	DF	WSD	L	YENWN	PGQDS	YFL	SGENC	VVM	WHD	QGQW	SDVPC	NHLY	STCK	MGL																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
MABA_RAT 1MSB/2MSB :														GKKFFVT	INHER	MPFS	KYKAL	CSLR	GTVA	IPR	NAE	ENKAI	QEVAKTS	AFLGIT	DEVTE	GGFM	YVTGG	RLT	YSNWK	KDEPN	DHSG	EDCV	TIVD	EDCV	ELYIK	REK	DVGM	WDER	CKSK	KLAL	CYTA																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																														
E-SELECTIN HUMAN IESL :														WSYNTS	TEAM	TYDE	ASAY	CCQR	YTH	LVAI	QNK	EEI	TEFL	NSLS	YS	SPSY	YWI	G	W	W	G	Y	W	P	C	I	W	C	L	C																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																															
Most frequent CRD residue (>70% conservation)														Y	W	A	C	G	L	I	E	F	L	WIGL	1	1	W	G	Y	W	P	C	I	W	C	L	C																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																		
Calcium sites 1 and 2																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																							

Figure 3 Residue conservation in 13 group I sequences of the CRD superfamily

A total of 73 residues conserved to more than 90% are indicated below the alignment. These are compared with 26 residues conserved to better than 70% in the 131 CRD sequences, the two Ca^{2+} -binding sites 1 and 2, and the observed secondary structures A1, A2 and B1–B7 found in the crystal structures of mannose-binding protein and E-selectin (A, α -helix; B, β -strand). Residue numbering follows that in Figure 2.

Figure 4 Molecular graphics modelling of the CRD

α -Helices and β -strands are identified as in Figures 2 and 3. All models are shown in the same orientation. The positions of the two disulphide bridges in each fold are denoted by thick black lines. (a) The G3 CRD model is shown with the conserved basic residues Arg-24, Arg-29, Arg-40, Arg-76 and Lys-133 marked by labelled spheres. The two α -helices and eight β -strands are labelled A1–A2 and B0–B7 respectively. (b) The PTR from TSG-6 (1tsg). This is oriented to show the similarity in structure with the CRD. The five conserved basic residues Arg-9/Lys-11 (on B1), Arg-56, Arg-63 and Arg-75 (on B5) are marked by unlabelled spheres in accordance with [23]. The secondary structure is labelled to follow (a). (c) The CRD from mannose-binding protein (2msb). Acidic residues at the two Ca^{2+} -binding sites 1 and 2 are marked by numbered spheres. The secondary structure is labelled to follow (a).

lower Z-scores for the long CRD sequences is best explained by the presence of only the short CRD fold in the basic set of 254 folds used by THREADER.

Molecular graphics modelling of the G3 CRD in human aggrecan

The sequence alignment between the G3 CRD of human aggrecan G3 and rat mannose-binding protein showed 25/129 identities (20%), and that between the G3 CRD and E-selectin showed 32/129 identities (26%). These identities fall within the 'twilight zone' of 15–30% for identifying structurally homologous protein folds [33,37,38]. The sequence, secondary structure and fold recognition analyses above, together with CD data [39], validate the homology modelling of the G3 CRD structure with the crystal structure of mannose-binding protein. In the G3 CRD homology model, 103 residues were set as invariant core residues,

whereas two surface loops with 10 residues and the extra N-terminal 14 residues were remodelled (see the Materials and methods section). Based on the recent determination of crystal structure of lithostathine [20], the N-terminal 14 residues were constructed as an extended β -strand (denoted B0 for numerical consistency), which was then remodelled to form an antiparallel β -sheet (B0–B1–B7–B2) with the appropriate connection for the disulphide bridge at Cys-11–Cys-22. Even though the model will not substitute for the detail of a crystal structure, it is precise enough for an appraisal of the conserved basic residues in the G3 CRD.

The topology of the aggrecan G3 CRD model is compared with the crystal structure of mannose-binding protein (2msb) and the NMR structure of the PTR from TSG-6 (1tsg) in Figure 4 [16,25]. The structural similarity of the $2\alpha/3\beta$ motif in both the CRD and PTR folds is clear, whereas the upper region of the two

Figure 5 Comparison of the G3 SCR sequences with those of SCR-15 and SCR-16 in human factor H

The G3 sequences are identified by their SWISSPROT or PIR accession names or numbers. Residues conserved in more than 90% of the G3 SCR sequences are indicated below the alignment. These are compared with the 22 best-conserved residues in an alignment of 101 SCR sequences ([28], Table 1). The SCR structures are identified by their Brookhaven codes (1hfi, 1hcc and 1hfh). Secondary structure analyses by DSSP follow the abbreviations given in Methods. The six β -strands from the averaged structure predictions [28] and the averaged DSSP analyses are indicated by B1 to B6. Side-chain accessibility analyses by COMPARER follow those in Methods, and are compared with averaged hydropathy predictions [28] (e, exposed; b, buried).

folds is different [23]. Four of the five conserved basic residues in the G3 CRD model (Arg/Lys-24, Arg-29, Arg-40 and Arg/Lys-76) are located on the front face of the CRD as viewed in Figure 4 (B1, B1, A1, and after B3, in that order). The fifth conserved basic residue (Lys-133) is at the rear of the CRD as viewed. The proximity of the first four residues in this model supports the significance of their 100% conservation in Figure 3. The conserved basic residues in the PTR (Arg-9/Lys-11, Arg-56, Arg-63 and Arg-75 [23]) are located at the top of the PTR in Figure 4 (B1, before B5, B5, and after B5, in that order). Arg-29 in the G3 CRD model occupies a similar topological position to Lys-11 on B1 in the PTR, as predicted from the sequence alignment. The proximity of these conserved basic residues on one surface of each of the G3 CRD and the PTR folds implies that these residues are functionally significant. Because different surfaces on the two domains are involved, this implies that the binding specificity of the G3 CRD is not for hyaluronate, which binds to the PTR.

The location of the five conserved basic residues relative to that of nine acidic residues that form the two putative Ca^{2+} -binding sites in the G3 CRD is of interest. The basic residues are located at the opposite end of the CRD to the acidic residues labelled 1 and 2 that bind carbohydrate (Figures 4a and 4c). Even though Arg/Lys-76 is adjacent to Asp-75 at Ca^{2+} site 1, it can be seen that the basic and acidic sets of residues occupy different areas of the G3 CRD surface. Biochemical data show that the G3 CRD exhibits Ca^{2+} -dependent C-type lectin activities [19,39–42]. The sugar specificity of aggrecan G3 CRD is for

galactose or fucose, not mannose or *N*-acetylglucosamine [39,40]. Despite the specificity for galactose that is predicted from the presence of Gln-99–Asp-101, galactose binds only weakly to rat aggrecan G3 CRD. Chicken versican G3 CRD is, however, specific for mannose and galactose [41], whereas human versican G3 CRD is specific for fucose or *N*-acetylglucosamine [42]. Crystallographic studies have shown that the preferential binding of sugars is dependent on the equatorial/axial or equatorial/equatorial arrangement of the 3-/4-hydroxy groups. This specificity is determined by the stereochemistry of the five side chains that ligate Ca^{2+} at site 2 [36,43–45]. The presence of Gln-99–Asp-101 at Ca^{2+} site 2 in G3 instead of Glu-99–Asn-101 accounts for the galactose specificity of the aggrecan CRDs but does not explain the different affinities of the versican CRDs. Together with the replacement of Asn-102 at Ca^{2+} site 1 by Ser-102 in brevican and versican, as well as the presence of hydrophobic residues near Ca^{2+} site 2, both biochemical and sequence data suggest that the G3 CRDs display a range of carbohydrate specificities in different members of this group. However, the presence of the same conserved basic residues in the G3 CRDs suggests that this domain has a similar functional role in proteoglycans.

Molecular modelling of the G3 SCR in human aggrecan

The alignment of 13 G3 SCR sequences (Figure 5) shows that the consensus length is 62 residues with no gaps or insertions and that residue conservation is high, with 30/62 residues that are

Figure 6 Molecular model of the CRD and SCR domains in G3 of human aggrecan

The relative orientation of the two domains is not known. (a) The two domains in G3 are linked by an extended polypeptide strand. The side-chain structures of 10 conserved basic residues are shown in full, and secondary structures are labelled as in Figures 3 and 5. The two disulphide bridges in each of the CRD and SCR domains are denoted by thick black lines. (b, c, d) Electrostatic potential maps for the G3 model. Starting from the orientation shown in (a), the three views [(b), (c) and (d), in that order] correspond to successive 90° rotational steps about the vertical axis. Red, negative charge; blue, positive charge.

conserved or conservatively replaced at greater than 90%. Full conservation is observed for all the cysteine residues (Cys-5, Cys-34, Cys-48 and Cys-61), six hydrophobic residues (Val-3, Tyr-14, Tyr-32, Ile-46, Trp-54 and Ile-59), three proline residues (Pro-8, Pro-9 and Pro-57), four glycine residues (Gly-1, Gly-6, Gly-18 and Gly-37) and two arginine residues (Arg-31 and Arg-47). The previous analysis of an alignment of 101 SCR sequences showed that 22/61 residue positions were well conserved [28], of which there are 15 matches with the G3 SCR sequences (Figure 5). These matches are all hydrophobic and show that the G3 SCR has a conserved hydrophobic core. Interestingly, Figure 5 showed that the G3 SCR sequences are atypical of the SCR superfamily in that they contained five highly conserved basic residues (Arg/Lys-21, Arg/Lys-23, Arg-31, Arg/His-41 and Arg-47) that are not conserved in other SCRs. In the 101 SCR sequences in [28], only 3 of these 5 residues occur in 5 sequences, 2 in 20 sequences, 1 in 39 sequences, and 0 in 37 sequences, which is as expected on the basis of random events. The Lys and Arg residues comprise 17% of the residues in the G3 SCR sequences. The C-terminal sequences after the G3 SCR also have a high proportion of basic residues (Figure 5). In Figure 5 these constitute 457 residues, of which 116 (25%) are Lys or Arg, whereas a typical protein would contain 5.7% Lys and 5.7% Arg residues [37].

Because all the SCR sequences formed one alignment, and there are no sequence subgroups as found for the CRD superfamily, a G3 SCR homology model could be constructed by using a structure derived from complement factor H, which was the closest to the G3 SCR in length (1hfh). An SCR structure contains six β -strands that form three antiparallel pairs (B1 and B2, B3 and B6, B4 and B5) [28,29]. The G3 SCR model was based on a conserved core structure of 48 β -strand and loop residues; 14 residues at three surface regions and the N- and C-termini were remodelled with little perturbation of the starting structure. All five conserved basic residues in G3 SCR were found on one extended surface on the G3 SCR model (Figure 6a). Lys-21/Arg-23 occurred near the junction of the SCR with

the CRD in a loop region, Arg-31/Arg-47 were located at the centre of the SCR (β -strands B2 and B4), and Arg/His-41 was located at the C-terminal region of the SCR (β -strand B3).

The G3 CRD and SCR models were joined together. The use of five secondary structure predictions with the 13 G3 SCR sequences showed that a β -strand was well predicted for Thr-2–Ala-4 in Figure 5. As this represents a continuation of the CRD β -strand B7 (Figure 3), and as the DSSP analyses of the SCR NMR structures (Figure 5) showed some β -strand in this region, the CRD and SCR domains in G3 were accordingly joined by an extended polypeptide link. This was long enough to allow the connection to be made, even though the rotation about this link would be arbitrary. A similar link joined the CRD and EGF domains in the crystal structure of E-selectin [17]. The dimensions of the G3 model were 7.0 nm $\times$ 5.7 nm $\times$ 3.9 nm. These compare well with the diameter of 8.3 ± 1.3 nm reported for chicken G3 after rotary shadowing [4], and the range of diameters of 8–11 nm (± 2 nm) similarly reported for several types of bovine G3 [46]. (Note that the decoration effect in rotary shadowing leads to a systematic overestimation of 2–3 nm in diameter.) If an EGF domain (1pfx) were present in G3, the length of G3 would be increased maximally by 3 nm. This might account for the higher diameters observed for G3 from sclera and tendon proteoglycans [46]. If the SCR domain is absent from G3, the resulting G3 CRD model has dimensions of 4.9 nm $\times$ 4.7 nm $\times$ 4.1 nm. This might account for the smaller diameter of 3.5 ± 1 nm for the G3 domain reported in [3].

Electrostatic surface calculations of the G3 CRD and SCR domains

The conserved basic residues in the G3 CRD model might be important for forming contacts with anionic substrates. Basic residues constitute the basis for link protein–hyaluronate interactions via the PTR domains [47,48], even though protein–carbohydrate interactions might be mediated also by hydrogen bond and hydrophobic contacts [49]. First, electrostatic surface

potentials were calculated for mannose-binding protein and E-selectin as a control. Both proteins generally exhibited relatively neutral surfaces with isolated small areas of positive and negative charge with the exception of a large region of negative charge at the top of the CRD fold, which corresponds to the two Ca^{2+} -binding sites 1 and 2 in mannose-binding protein (Figure 4c) and a smaller region in E-selectin that corresponds to the single Ca^{2+} -binding site 2. In application to the G3 CRD model, a large negative charge was seen at the region of the Ca^{2+} -binding sites 1 and 2 (Figure 6) that was similar in size to that seen for mannose-binding protein. In marked contrast with the control calculations, the G3 CRD model now exhibited a large area of positive charge close to the conserved residue Arg-24, together with two more close to the conserved residues Arg-29 and Arg-76 (Figure 4a). The distance between the α -carbon atoms of Arg-24 and Arg-29 was 1.4 nm, and that between Arg-29 and Arg-76 was similarly 2.2 nm. These distances are comparable to the separation of carboxy or sulphate groups in anionic polysaccharides such as hyaluronate, chondroitin sulphate or heparin (1hya, 2hya, 3hya, 4hya, 1c4s, 2c4s, 1hpn).

Similar calculations were performed for the G3 SCR domain. Control calculations for the SCR structures from complement factor H (Figure 5) showed an even distribution of positive and negative charges. Again in marked contrast with this, the G3 SCR model exhibited three large areas of positive charge on one side of the β -sheet sandwich, whereas the other side was negatively charged. These positive surfaces coincided with Lys-21/Arg-23, Arg-31/Arg-47 and Arg-41 (Figure 5). The α -carbon separations are 1.6 ± 0.2 nm between Lys-21/Arg-23 and Arg-31/Arg-47, and 1.5 ± 0.2 nm between Arg-31/Arg-47 and Arg-41. Even though no information is available on the interdomain orientation of the CRD and SCR models, the combination of both models in Figure 6 shows prominent areas of basic charge. In the interdomain orientation shown in Figure 6, they form a cleft lined by positive charges. The inclusion of Ca^{2+} ions in the G3 CRD in the surface potential calculations will diminish the magnitude of the acidic region at the top of Figure 6. The G3 model might be consistent with the proposed binding of G3 to a polyanionic ligand.

Conclusions

Structure, function and evolution of G3 in proteoglycans

This study has reported novel aspects of the CRD domain that were previously unnoticed and can now be subjected to experimental tests, and clarifies the relationship between the N-terminal and C-terminal domains in the proteoglycan family and with link protein. The molecular modelling of the G3 CRD and SCR domains unexpectedly revealed two surface areas of conserved positive charge in all members of the proteoglycan G3 family. The conservation of basic residues in proximity to each other raises the possibility that the G3 CRDs might interact with anionic polysaccharides, and there is evidence for this [41,50]. It is unlikely that G3 binds to hyaluronate because G1 already exists with this function. Similarly, it is unlikely that G3 binds to chondroitin sulphate or keratan sulphate because large amounts of these are found in aggrecan. If G3 has evolved to bind to an anionic ligand, perhaps to a sulphated one such as heparan sulphate or to a sialylated oligosaccharide, these interactions might be related to the secretion of aggrecan [8]. Heparin binds to G3 from versican, which contains both a CRD and a SCR domain at a site that is different from that for monosaccharides; removal of the SCR domain abolished G3 binding to heparin or heparan sulphate [41]. Our observation of conserved basic residues in G3 is consistent with known protein interaction sites

for anionic polysaccharides. Those involving heparin have been described for basic fibroblast growth factor and antithrombin III, both of which involve an arrangement of four or five basic residues [51,52]. In contrast, a survey of known lectin-carbohydrate interactions shows that acidic and amide groups commonly form the basis of hydrogen bond formation between protein and carbohydrate hydroxy groups in complex-type and high-mannose-type oligosaccharides [49]. Although Arg-97 and Lys-113 have been implicated in E-selectin adhesion to neutrophils [17], Arg-97 in that study corresponds to the loop between residues 115 and 116 in Figure 2 and is not conserved in P-selectin, whereas Lys-113 corresponds to residue 128 in Figure 2 and is conserved in the selectins and some of the group III CRDs. Neither Arg-97 or Lys-113 corresponds in position to the five conserved basic residues observed in the G3 CRDs (Figure 3). The present G3 modelling of conserved basic residues will assist the rational planning of experiments designed to explore G3 function in more detail.

The molecular modelling of G3 also indicates that G3 is more similar to G1 and G2 in structure than was previously believed. The two PTR domains in G1 form a very stable interaction with hyaluronate in conjunction with two more PTR domains in link protein. Protein structure analyses and NMR spectroscopy showed that the PTR contains a CRD-like protein fold [23,25]. In addition to the occurrence of CRD-like structures at both the N-terminal and C-terminal ends of aggrecan, both the PTR and G3 CRD exhibit conserved basic residues that are clustered on one surface of both domains. One of these (Arg-9/Lys-11) lies in the same topological position in both the G3 CRD and the PTR domains. This residue has been implicated in the binding of hyaluronate to the PTR [35]. The exon structures of the G3 CRD and the PTR are also related in a manner that is related to their similar structures [23]. These observations suggest a evolutionary relationship between the PTR and CRD domains.

In relation to carbohydrate binding at the two putative Ca^{2+} -binding sites in the G3 CRD, the affinities of the G3 CRD domains for carbohydrate are almost an order of magnitude weaker than those of other CRD groups [39]. No clear indications of Ca^{2+} -dependent carbohydrate binding were obtained from the present modelling, as the specificity of the G3 CRD varies from proteoglycan to proteoglycan, and similarly, the residues that are adjacent to the Ca^{2+} -binding sites. Recently the CRD in G3 of versican has been found to bind to a possible sugar moiety on the glycoprotein tenascin-R in a Ca^{2+} -dependent manner, where tenascin-R is a cell adhesion protein in the granular layer of rat cerebellum [42].

Fold recognition and molecular modelling of the CRD domain

The present molecular modelling of the G3 CRD was based jointly on an alignment of 131 CRD sequences, consensus secondary structure predictions and averaged protein fold recognition analyses, unlike earlier modelling studies of several long CRD domains [30–32]. The consensus secondary structure prediction for 131 CRD sequences was 82% accurate on a residue-by-residue basis when compared with the observed secondary structure in the crystal structures, which is higher than usually found [33]. It failed to detect the β -strand B6, which is attributable to the presence of residues required for the conserved Ca^{2+} site 2 on B6. The predictions were useful in highlighting the existence of the β -strand B2 that turned out to be present in the CRD crystal structures. The averaged protein fold recognition analyses worked well for the short CRD sequences, and slightly less well for the long CRD sequences. Protein prediction analyses were

able to analyse the CRD structure successfully, starting from a sequence alignment for this superfamily.

Fold recognition of the PTR domain

In the course of this study and that in [23], the NMR structure of the PTR domain from a cell-surface proteoglycan TSG-6 was determined [25] (1tsg). We had predicted that the PTR structure was structurally related to the CRD, and now comment on this. The consensus secondary structure predictions for the PTR in [24] and [23] are 81% and 85% accurate respectively on a residue-by-residue basis according to the assignments in the NMR structure [25]. Interestingly, the mispredicted CRD β -strand B6 is topologically equivalent to B5 in the PTR; this was now predicted correctly. This is attributable to the absence of the residues responsible for the CRD Ca^{2+} site. However, the PTR prediction missed the short β -strand B3. The averaged protein fold recognition analyses correctly correlated the CRD and PTR folds. The PTR prediction did not detect a realignment of the α -helix A2 in the opposite direction to that seen in the CRD (Figures 4a and 4b). This successful outcome of the PTR predictive analysis is very similar to that for our prediction of the von Willebrand factor Type A structure in the absence of its crystal structure [33,34].

We thank Professor T. E. Hardingham for his interest and useful discussions. N.C.B. thanks the B.B.S.R.C. for a CASE studentship, and Eli Lilly for support. S.J.P. thanks the Wellcome Trust and the Royal Society for other support.

REFERENCES

- Hardingham, T. E. and Fosang, A. J. (1992) *FASEB J.* **6**, 861–870
- Hardingham, T. E., Fosang, A. J. and Dudhia, J. (1992) In *Articular Cartilage and Osteoarthritis* (Kuettner, K., Schleyerbach, R., Hascall, V. C. and Peyron, J. G., eds.), pp. 5–20, Raven Press, New York
- Paulsson, M., Mörgelin, M., Wiedemann, H., Beardmore-Gray, M., Dunham, D., Hardingham, T. E., Heinegård, D., Timpl, R. and Engel, J. (1987) *Biochem. J.* **245**, 763–772
- Dennis, J. E., Carrino, D. A., Schwartz, N. B. and Caplan, A. I. (1990) *J. Biol. Chem.* **265**, 12098–12103
- Dudhia, J., Davidson, C. M., Wells, T. M., Vynios, D. H., Hardingham, T. E. and Bayliss, M. T. (1996) *Biochem. J.* **313**, 933–940
- Li, H., Schwartz, N. B. and Vertel, B. M. (1993) *J. Biol. Chem.* **268**, 23504–23511
- Vertel, B. M., Walters, L. M., Grier, B., Maine, N. and Goetinck, P. F. (1993) *J. Cell Sci.* **104**, 939–948
- Luo, W., Kuwada, T. S., Chandrasekaran, L., Zheng, J. and Tanzer, M. L. (1996) *J. Biol. Chem.* **271**, 16447–16450
- Baldwin, C. T., Reginato, A. M. and Prockop, D. J. (1989) *J. Biol. Chem.* **264**, 15747–15750
- Doege, K. J., Sasaki, M., Horigan, E., Hassell, J. R. and Yamada, Y. (1991) *J. Biol. Chem.* **266**, 894–902
- Fülöp, C., Walcz, E., Vallyon, M. and Giant, T. T. (1993) *J. Biol. Chem.* **268**, 17377–17383
- Drickamer, K. (1988) *J. Biol. Chem.* **263**, 9557–9560
- Drickamer, K. (1993) *Curr. Opin. Struct. Biol.* **3**, 393–400
- Drickamer, K. (1993) *Prog. Nucleic Acid Res. Mol. Biol.* **45**, 207–232
- Weis, W. I., Kahn, R., Fourme, R., Drickamer, K. and Hendrickson, W. A. (1991) *Science* **254**, 1608–1615
- Weis, W. I., Drickamer, K. and Hendrickson, W. A. (1992) *Nature (London)* **360**, 127–134
- Graves, B. J., Crowther, R. L., Chandran, C., Rumbergwer, J. M., Li, S., Huang, K., Presky, D. H., Familletti, P. C., Wollitzky, B. A. and Burns, D. K. (1994) *Nature (London)* **367**, 532–538
- Weis, W. I. (1994) *Structure* **2**, 147–150
- Ng, K. K.-S., Drickamer, K. and Weis, W. I. (1996) *J. Biol. Chem.* **271**, 663–674
- Bertrand, J. A., Pignol, D., Bernard, J.-P., Verdier, J.-M., Dagorn, J.-C. and Fontecilla-Camps, J. C. (1996) *EMBO J.* **15**, 2678–2684
- Larsen, I. K., Nielsen, B. B., Rasmussen, H. and Kastrup, J. S. (1996) XVIIth Congress of the International Union of Crystallography, Seattle, Washington, U.S.A. Abstract PS04.18.06
- Brissett, N. C. and Perkins, S. J. (1996) *Biochem. Soc. Trans.* **24**, 99S–99S
- Brissett, N. C. and Perkins, S. J. (1996) *FEBS Lett.* **388**, 211–216
- Perkins, S. J., Nealis, A. S., Dudhia, J. and Hardingham, T. E. (1989) *J. Mol. Biol.* **206**, 737–754
- Kohda, D., Morton, C. J., Parkar, A. A., Hatanaka, H., Inagaki, F. M., Campbell, I. D. and Day, A. J. (1996) *Cell* **86**, 767–775
- Norman, D. G., Barlow, P. N., Baron, M., Day, A. J., Sim, R. B. and Campbell, I. D. (1991) *J. Mol. Biol.* **219**, 717–725
- Barlow, P. N., Steinkasserer, A., Norman, D. G., Kieffer, B., Wiles, A. P., Sim, R. B. and Campbell, I. D. (1993) *J. Mol. Biol.* **232**, 268–284
- Perkins, S. J., Haris, P. I., Sim, R. B. and Chapman, D. (1988) *Biochemistry* **27**, 4004–4012
- Molina, H., Perkins, S. J., Guthridge, J., Gorka, J., Kinoshita, T. and Holers, V. M. (1995) *J. Immunol.* **154**, 5426–5435
- Padlan, E. A. and Helm, B. (1993) *Receptor* **3**, 325–341
- Bajorath, J. and Aruffo, A. (1994) *J. Biol. Chem.* **269**, 32457–32463
- Sönnichsen, F. D., Sykes, B. D. and Davies, P. L. (1995) *Protein Sci.* **4**, 460–471
- Sternberg, M. J. E. (1997) *Protein Structure Prediction*, IRL Press, Oxford
- Edwards, Y. J. K. and Perkins, S. J. (1996) *J. Mol. Biol.* **260**, 277–285
- Peach, R. J., Hollenbaugh, D., Stamenkovic, I. and Aruffo, A. (1993) *J. Cell Biol.* **122**, 257–264
- Kolatkar, A. R. and Weis, W. I. (1996) *J. Biol. Chem.* **271**, 6679–6685
- Creighton, T. E. (1993) *Proteins: Structures and Molecular Properties*, Freeman, New York
- Taylor, W. R. (1988) *Protein Eng.* **2**, 77–86
- Saleque, S., Ruiz, N. and Drickamer, K. (1993) *Glycobiology* **3**, 185–190
- Halberg, D. H., Proulx, G., Doege, K., Yamada, Y. and Drickamer, K. (1988) *J. Biol. Chem.* **263**, 9486–9490
- Ujita, M., Shinomura, T., Ito, K., Kitagawa, Y. and Kimata, K. (1994) *J. Biol. Chem.* **269**, 27603–27609
- Aspberg, A., Binkert, C. and Ruoslahti, E. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 10590–10594
- Drickamer, K. (1992) *Nature (London)* **360**, 183–186
- Iobst, S. T. and Drickamer, K. (1994) *J. Biol. Chem.* **269**, 15512–15519
- Blanck, O., Iobst, S. T., Gabel, C. and Drickamer, K. (1996) *J. Biol. Chem.* **271**, 7289–7292
- Mörgelin, M., Paulsson, M., Malmström, A. and Heinegård, D. (1989) *J. Biol. Chem.* **264**, 12080–12090
- Hardingham, T. E., Ewins, R. J. F. and Muir, H. (1976) *Biochem. J.* **157**, 127–143
- Lyon, M. (1986) *Biochim. Biophys. Acta* **881**, 22–29
- Weis, W. I. and Drickamer, K. (1996) *Annu. Rev. Biochem.* **65**, 441–473
- Yang, B., Yang, B. L. and Goetinck, P. F. (1995) *Anal. Biochem.* **228**, 299–306
- Faham, S., Hileman, R. E., Fromm, J. R., Linhardt, R. J. and Rees, D. C. (1996) *Science* **271**, 1116–1120
- Kridel, S. J. and Knauer, D. J. (1997) *J. Biol. Chem.* **272**, 7656–7660