The CUB Domain

A Widespread Module in Developmentally Regulated Proteins

Peer Bork[1,2] and Georg Beckmann[2]

[1]*EMBL, 6900 Heidelberg, Germany*
[2]*Max-Delbrück-Centre for Molecular Medicine*
1115 Berlin-Buch, Germany

(Received 27 October 1992; accepted 5 February 1993)

Sequence analysis has revealed the presence of 31 copies of an extracellular domain, here called CUB, in 16 functionally diverse proteins such as the dorso-ventral patterning protein tolloid, bone morphogenetic protein 1, a family of spermadhesins, complement subcomponents Cls/Clr and the neuronal recognition molecule A5. Most of them are known to be involved in developmental processes. Our analysis of this new family includes the identification of seven previously undescribed members, the characterization of conserved features and a topology prediction of this approximately 110 residue spanning domain, which suggests an antiparallel β-barrel similar to those in immunoglobulins.

Keywords: development; modules; sequence homology; superfamily; topology

Communication between cells during development requires a network of defined interactions. Nature has apparently solved this problem by combining structurally and functionally independent domains (modules), which are sometimes the only link between otherwise distinct proteins (for a review, see Bork, 1992, and references therein).

One out of many examples is a module that was first identified in the complement subcomponents Cls and Clr (Cls/Clr†), in both of which it occurs twice (Leytus *et al.*, 1986; Tosi *et al.*, 1987). Later on, the domain was found in an embryonic sea-urchin protein (Uegf, or fibropellin; Delgadillo-Rosso *et al.*, 1989) and a relationship has been shown to bone morphogenetic protein 1 (Bmpl), to a calcium dependent serine protease (Casp) as well as to the embryonic protein Uvs2 from *Xenopus laevis* (Bork, 1991). The latter two belong to a family of proteins that also contain a domain with zinc-metalloprotease activity (for a review of this protease, see PROSITE database, accession number PS00142; Bairoch, 1992). Recently, several other proteins containing a zinc-metalloprotease as well as domains related to Clr/Cls have also been reported: the *Drosophila* dorso-ventral patterning gene product tolloid (Shimell *et al.*, 1991), the blastulla restricted proteins Bp10 (Lepage *et al.*, 1992) and Span (Reynolds *et al.*, 1992) both from sea urchin. Based on the first three identified proteins of this family we propose the name CUB (**C**omplement subcomponents Clr/Cls, **U**egf, **B**mpl) for this widespread domain, which has now also been found in the neuronal recognition molecule A5 (Takagi *et al.*, 1991) and the tumor necrosis factor-induced protein Tsg6 (Lee *et al.*, 1992).

Here we summarize these similarities (Fig. 1, Fig. 2) and describe the detection of several new members: (1) the zona pellucida-binding spermadhesins (Awn, Aqnl, Aqn3; Sanz *et al.*, 1991, 1992*a*,*b*), (2) acidic seminal fluid protein described as a new growth factor (Asfp; Wempe *et al.*, 1992) as well as (3) the astroglial (choroid plexus) protein p14 (p14; Lecain *et al.*, 1991). In addition, the DNA next to a novel serum inducible gene in vascular smooth muscle cells (Sig; L. Liau and P. Feng, unpublished results; EMBL accession number M86381) revealed 95% sequence identity to human *Tsg6* and might be its orthologous gene in rabbit.

The newly described members have been identi-

† Abbreviations used: Clr/Cls, complement subcomponents Clr/Cls; Uegf, epidermal growth factor related sea urchin protein; Bmpl, bone morphogenetic protein 1; Casp, calcium dependent serine protease; Uvs2, developmentally regulated protein UVS.2; Bp10, blastula protein 10; Span, very early blastula genproduct Span; *Tsg6*, product of the TNF stimulated gene; TNF, tumor necrosis factor; Awn, Aqnl, Aqn3, spermadhesins; Asfp, acidic seminal fluid protein; p14, choroid plexus protein p14; Sig, serum induced glycoprotein.

0022–2836/93/100539–07 $08.00/0

Table 1
Pairwise percentage identities of the newly detected CUBs to all other known members of this family

	p14-1	p14-2	Asfp	Awn	Aqn1	Aqn3
p14-1	# #					
p14-2	**35**	# #				
Asfp	27	25	# #			
Awn	28	21	**44**	# #		
Aqn1	25	24	**47**	**58**	# #	
Aqn3	26	23	**44**	**46**	**53**	# #
A5-1	**39**	**42**	**33**	27	**30**	22
A5-2	**30**	**30**	23	26	19	22
TSG-6	**35**	**36**	25	**34**	29	27
Sig	**35**	**37**	25	**32**	26	26
Uegf†	19	26	17	18	17	12
Bmp1-1	**39**	**33**	23	26	28	28
Bmp1-2	**38**	**34**	29	29	25	27
Bmp1-3	**34**	**42**	25	25	27	29
Tld-1	26	19	25	24	27	**30**
Tld-2	**35**	**41**	28	29	25	23
Tld-3	27	**32**	24	25	23	24
Tld-4	**30**	**35**	20	24	21	21
Tld-5	**30**	28	21	24	25	19
Uvs-1	**31**	**32**	19	22	20	23
Uvs-2	**34**	**33**	**30**	29	**32**	**30**
Bp10-1	**36**	**38**	24	21	27	28
Bp10-2	**34**	**37**	29	25	21	18
Span-1	**32**	**37**	25	22	29	28
Span-2	**36**	**32**	27	24	21	27
C1r-1	**37**	**33**	22	15	19	15
C1s-1	**31**	**36**	24	20	20	18
Casp-1	29	**36**	27	27	26	19
C1r-2	**39**	27	17	18	19	15
C1s-2	**32**	26	15	21	23	18
Casp-2	28	28	14	21	29	19
The two	260	270	141	128	120	111
best	Bmp1-1	Bmp1-3	Bmp1-2	Bmp1-2	Bmp1-3	Sig
FASTA	260	254	121	117	111	110
hits	A5-1	Bp10-1	Tld-2	Sig	Bmp1-2	Tld-1

A complete dynamic programming alignment was used as implemented in the BESTFIT program of the GCG package (Devereux *et al.*, 1984). Numbers are underlined if the percentage identity is above the threshold for structural homology (Sander & Schneider, 1991). They are shown in bold if they exceed 30%. The optimized scores of the 2 highest ranking proteins (neglecting scores of the query sequences) in a FASTA search (Pearson & Lipmann, 1988) are also shown.

† Only a part of the domain in Uegf has been sequenced.

fied by a consensus property pattern constructed for this domain (Bork, 1991). All domains of the CUB family scored better than the random background of unrelated sequences (for details of this approach, see Bork & Grunwald, 1990; Bork, 1991). When attempting to establish remote relationships, a significance estimate should be provided. Since there are many insertions/deletions in the multiple alignment (Fig. 1) that are not yet included in mathematical treatments of significance estimates, we verified our findings by the use of standard sequence analysis methods as follows. (1) FASTA (TFASTA) homology searches (Pearson & Lipmann, 1988) with the candidate domains were carried out and the output was sorted for the optimized scores, which include weights for similar amino acids and gap penalties. In all cases, several regions of known CUBs had optimized scores above 110, which are at least indicative of a homology (Table 1). Since FASTA often detects only parts of domains due to insertion/deletions, we used (2) a complete dynamic programming algorithm for pairwise alignments as implemented in the GCG package (Devereaux *et al.*, 1984). These comparisons resulted in high pairwise identities between the newly detected CUBs and known members of this family (Table 1). Often, they have more than 25% amino acid identity over the whole domain, above the threshold for structural homology for globular proteins (Sander & Schneider, 1991). In addition, pairwise identities to some of the known CUBs above 30% (Table 1) leave little doubt about the relationships. Finally, (3) in a PROFILESEARCH (Gribskov *et al.*, 1987) with the alignment in Figure 1, all CUBs had higher scores than all other proteins in current sequence databases.

Interestingly, as known for nearly all of the extracellular modules (Bork & Doolittle, 1992), we were not able to identify any CUB in prokaryotes, plants or yeast. One might argue that this is due to

```
                    1            2                     3            4    5
β-strands:         bbbb        bbbbbb               bbbbbb       bbbbb bbbb
consensus:       Ct  h t     t h ttthtt           tC a I ht t  h h htth ht
Awn_Pig        9 CGGVLRDP...PGKIFNSDGPQ..........KDCVWTIKVKPH..FHVVLAIPPLNLS.......
Aqn1_Pig       9 CGGVLRNY...SGRISTYEGPK..........TDCIWTILAKPG..SRVFVAIPYLNLA.......
Aqn3_Pig       9 CGGFLKNY...SGWISYYKALT..........TNCVWTIEMKPG..HKIILQILPLNLT.......
Asfp/Bovin    10 CGGILKEE...SGVIATYYGPK..........TNCVWTIQMPPE..YHVRVSIQYLQLN.......
Casp_Mesau-1  17 ASFSAEPTM..HGEILSPNYPQ.AYPNE....MEKTWDIEVPEG..FGVRLYFTHLDMELSEN...
Co1s_Human-1  11 AWVYAEPTM..YGEILSPNYPQ.AYPSE....VEKSWDIEVPEG..YGIHLYFTHLDIELSEN...
Co1r_Human-1  17 GSIPIPQKL..FGEVTSPLFPK.PYPNN....FETTTVITVPTG..YRVKLVFQQFDLEPSEG...
Tld_Drome-3  624 CGGVVDATK.SNGSLYSPSYPD.VYPNS....KQCVWEVVAPPN..HAVFLNFSDLEGTRFHY<2>
Tld_Drome-4  787 CKFEITTS...YGVLQSPNYPE.DYPRN....IYCYWHFQTVLG..HRIQLTFHDFEVESHQE...
A5p/Xenla-2  147 CSRNFTSS...NGVIKSPKYPE.KYPNA....LECTYIIFAPKM..QEIVLEFESFELEADSN<6>
Bmp1_Human-2 435 CGGDVKKD...YGHIQSPNYPD.DYRPS....KVCIWRIQVSEG..FHVGLTFQSFEIERHDS...
Tld_Drome-2  468 CGGDLKLTK..DQSIDSPNYPM.DYMPD....KECVWRITAPDN..HQVALKFQSFELEKHDG...
Bmp1_Human-3 591 CGGFLTKL...NGSITSPGWPK.EYPPN....KNCIWQLVAPTQ..YRISLQFDFFETEGNDV...
Bmp1_Human-1 322 CGETLQDS...TGNFSSPEYPN.GYSAH....MHCVWRISVTPG..EKIILNFTSLDLYRSRL...
A5p/Xenla-1   27 CGDTIKITS..PSYLTSAGYPH.SYPPS....QRCEWLIQAPEHY.QRIMINFNPHFDLEDRE...
P14/Mouse-1   36 CGGDVTGE...SGYVASEGFPN.LYPPN....KKCIWTITVPEG..QTVSLSFRVFDMELHPS...
P14/Mouse-2  158 CGGRMEKA...QGTLTTPNWPESYYPPG....ISCSWHIIAPSN..QVIMLTFGKFDVEPDTY...
Sig/Rabit      ? CGGVFTDP...KRIFKSPGFPN.EYDDN....QICYWHIRLKYG..QRIHLSFLNFDLEYDPG...
Tsg6/Human   135 CGGVFTDP...KRIFKSPGFPN.EYEDN....QICYWHIRLKYG..QRIHLSFLDFDLEDDPG...
Pb10/Parli-1 339 CSERFTEM...TGVITSPNWPG.RYEDN....MACVYQIEGPPG..STIELTFTEMNIENHAA...
Span/Strpu-1 340 CSYRFTEM...TGEITSPNYPS.NYEDN....TACVYEIEGPYG..STIELTFLDMEIETETL...
Pb10/Parli-2 484 CGGSFGGT...QGRVATPNYPN.NYDND....LECVYVIEVEIG..RRVELDFIDFVLEDETN...
Span/Strpu-2 503 CGGTFVGV...EGRVASPNYPN.DYDNS....LQCDYVIEVDDG..RRVELIFEDFGLEDETT...
Uvs2/Xenla-1   ? CSNLLPYS...NGMMISANYPS.AYPNN....ANCVWLIRTPSG...QVTLQFQAFDIQSSSG...
Uvs2/Xenla-2   ? CGGAFYSS...PKTFTSPNYPG.NYTTN....TNCTWTITAPAG..FKVSLRITDFELEIGAS...
Tld_Drome-5  900 CGGYLRATNHSQTFYSHPRYGSRPYKRN....MYCDWRIQADPE..SSVKIRFLHFEIEYSER...
Co1s_Human-2 175 CSGDVFTAL..IGEIASPNYPK.PYPEN....SRCEYQIRLEKGFQVVVTLRREDFDVEAADS<3>
Casp_Mesau-2 181 CSGNVFTAL..IGEISSPNYPN.PYPEN....SRCEYQILLEEGFQVVVTIQREDFDVEPADS<3>
Co1r_Human-2 193 CSSELYTEA..SGYISSLEYPR.SYPPD....LRCNYSIRVERGLTLHLKF.LEPFDIDDHQQ<2>
Tld_Drome-1  330 CGRTYQQN...SGHIVSPHFI...YSGN<26>GNCEWITATNGE..KVILH.LQQLLMSSDD....
Uegf/Strpu    62 CGYNVFDA...NGMIDSPNYPA.MYNNR....ADCLYLVRITKA..RSITFTIEDFMTEVFKD>>>
                 +---------------------------------+

                  6              7                        8                 9
β-strands:     bbbbbbb         bbbbb                  bbbbbbb            bbbbbb
consensus:   C h-hhth tt       t h +hCGt              t h h a tD t t    tGF h a
Awn_Pig      CGKEYVELLDG..PPGSEIIGKICGG.ISLV....FRSSSNIATIKRLRTSGHRA...SPFHIYYYAD
Aqn1_Pig     CGKEYVEVQDG..LPGAGNYGKLCSG.IGLT....YQSSSNALSIKYSRTAGHSA...SSFDIYYYGD
Aqn3_Pig     CGKEYLEVRDQ..RAGPDNFLKVCGG.TGFV....YQSSHNVATVKYSRDSHHPA...SSFNVYFYGI
Asfp/Bovin   CNKESLEIIDG..LPGSPVLGKICEG.SLMD....YRSSGSIMTVKYIREPEHPA...SFYEVLYFQD
Casp_Mesau-1 CEYDSVQIISG.....GVEEGRLCGQRTSKNA<8>FQIPYNKLQVIFRSDFSNEE<2>TGFAAYYAAI
Co1s_Human-1 CAYDSVQIISG.....DTEEGRLCGQRSSNNP<8>FQVPYNKLQVIFKSDFSNEE<2>TGFAAYYVAT
Co1r_Human-1 CFYDYVKISAD.....KKSLGRFCGQ.LGSPL<8>FMSQGNKMLLTFHTDFSNEE<7>KGFLAYYQAV
Tld_Drome-3  CNYDYLIIYSKMRDNRLKKIGIYCGHELPPVV...NSE.QSILRLEFYSDRTVQR...SGFVAKFVID
Tld_Drome-4  CIYDYVAIYDG.RSENSSTLGIYCGGREPYAV...IAS.TNEMFMVLATDAGLQR...KGFKATFVSE
A5p/Xenla-2  CRYDWLGIWDG.FPGVGPHIGRYCGQNTPGRV...RSF.TGILSMIFHTDSAIAK...EGFFANFSVV
Bmp1_Human-2 CAYDYLEVRDG.HSESSTLIGRYCGYEKPDDI...KST.SSRLWLKFVSDGSINK...AGFAVNFFKE
Tld_Drome-2  CAYDFVEIRDG.NHSDSRLIGRFCGDKLPPNI...KTR.SNQMYIRFVSDSSVQK...LGFSAALMLD
Bmp1_Human-3 CKYDFVEVRSG.LTADSKLHGKFCGSEKPEVI...TSQ.YNNMRVEFKSDNTVSK...KGFKAHFFSE
Bmp1_Human-1 CWYDYVEVRDG.FWRKAPLRGRFCGSKLPEPI...VST.DSRLWVEFRSSSNWVG...KGFFAVYEAI
A5p/Xenla-1  CKYDYVEVIDG.DNANGQLLGKYCGKIAPSPL...VST.GPSIFIRFVSDYETPG...AGFSIRYEVF
P14/Mouse-1  CRYDALEVFAG.SGTSGQRLGRFCGTFRPAPV...VAP.GNQVTLRMTTDEGTGG...RGFLLWYSGR
P14/Mouse-2  CRYDSVSVFNGAVSDDSKRLGKFCGDKAPSPI...SSE.GNELLVQFVSDLSVTA...DGFSASYRTL
Sig/Rabit    CLADYVEIYDS.YDDVHGFVGRYCGDELPEDI...IST.GNVMTLKFLSDASVTA...GGFQIKYVTV
Tsg6/Human   CLADYVEIYDS.YDDVHGFVGRYCGDELPDDI...IST.GNVMTLKFLSDASVTA...GGFQIKYVAM
Pb10/Parli-1 CRYDAVEVRKD...DINSDGEKFCGNTLPAVQ...ISS.GNQMLISFTSDPSITG...RGFRATYRIV
Span/Strpu-1 CRYDAVEVRKD...DINSIGEKFCGNTLPPVQ...ISS.SNQMMVSFTSDPSITR...RGFKATYVII
Pb10/Parli-2 CRWDSLSINLG...DGIKIDMKMCGREYPAAS...LVSIGNNMELTLISDRSVTD...RGFMADYRAI
Span/Strpu-2 CRWDSLMINLG...NGIKVGMKMCGREYPAAS...LVSIGNRMELKLKTDGSVND...RGFVASYRAI
Uvs2/Xenla-1 CVSDYIKIYDGPTKAFPVLVNRACGTGLIPLQ...IAS.TNQMLVEFVSDRAVTG...TGFKATYGSI
Uvs2/Xenla-2 CRYDYLNIYNS...TLGAVMGPYCGPIDFHSA...IVSKSNSMMITMNSDFSKQY...KGFSATYTFV
Tld_Drome-5  CDYDYLEITEE.GYSMNTIHGRFCGKHKPPII...ISN.SDTLLLRFQTDESNSL...RGFAISFMAV
Co1s_Human-2 CLDSLVFVAGD......RQFGPYCGHGFPGPL<1>IETKSNALDIIFQTDLTGQK...KGWKLRYHGD
Casp_Mesau-2 CQDSLLFAAKN......RQFGPFCGNGFPGPL<1>IETHSNTLDIVFQTDLTEQK...KGWKLRYHGD
Co1r_Human-2 CPYDQLQIYAN.....GKNIGEFCGKQRPPD....LDTSSNAVDLLFFTDESGDS...RGWKLRYTTE
Tld_Drome-1  CTQDYLEIRDG.YWHKSPLVRRICGNVSGEV....ITTQTSRMLLNYVNRNAAKG<1>RGFKARFEVV
             +-----------------------+
```

Fig. 1.

sampling bias, but it should be noted that in the whole yeast chromosome III none of the many well-characterized extracellular modules has been identified (Bork *et al.*, 1992).

A multiple alignment of all detected CUBs, which span about 110 amino acid residues, reveals the presence of several rather conserved blocks interrupted by variable regions of flexible length (Fig. 1). Four cysteine residues are conserved in all CUB modules except for the first in each of the complement subcomponents. These residues probably form two disulfide bridges (C1-C2; C3-C4; Fig. 1) as has been shown for Clr/Cls (Gagnon & Arlaud, 1985) and the spermadhesins: Aqnl (Sanz *et al.*, 1992*a*), Aqn3 (Sanz *et al.*, 1991) and Awn (Sanz *et al.*, 1992*b*). That Clr/Cls and the spermadhesins have the same arrangement of the disulfide bridges is a further indication of their relationship. In addition to various conserved hydrophobic and aromatic positions, only a few other residues are nearly invariant, which suggests that the different domains might have distinct binding specificities.

The conserved hydrophobicity patterns cover the whole domain and are typical for an antiparallel β-sheet. Secondary structure predictions also indicate an all-β structure. In particular, a new method, which has an average three-state accuracy of 70% (Rost & Sander, 1992; Rost & Sander, personal communication) predicts exactly the location of seven out of the nine strands proposed in Figure 1. The only inconsistency occurs in strand 2, where a relatively conserved proline had to be placed in a putative β-strand (Fig. 1). In spite of the fact that no protein with known three-dimensional structure appears to have sequence similarities to CUB, we propose an arrangement of the predicted antiparallel β-strands like that observed in the fold of immunoglobulin-like domains. This is supported by the location of the two putative disulfide bridges that would, respectively, connect the beginning of strand 1 with that of strand 3 and the beginning of strand 6 with the end of strand 7. Following this topology model, strand 6 and 7 could be further stabilized by a salt bridge made by the conserved negatively and positively charged positions in the corresponding strands (Fig. 1). Based on similar hydrophobicity patterns to those shown in Figure 1, Bazan (1990) proposed an immunoglobulin-like fold for cytokine receptors and fibronectin type III repeats. The recently determined three-dimensional structure for the latter (Main *et al.*, 1992, Leahy *et al.*, 1992; DeVos *et al.*, 1992) has proved the success of this topology prediction. The experimental and predicted structures only differed by a shift of strand 4 leading to a sheet-switch of this strand in the fibronectin type III domains. This has also been observed in the seven-stranded, immunoglobulin-like domains of CD4 and CD2 (Ryu *et al.*, 1990; Jones *et al.*, 1992). The immunoglobulin-like fold has also been found in numerous other modules with a length of about 90 to 110 amino acid residues, e.g. the two domains in PapD-like proteins (Holmgren *et al.*, 1992), a domain of several cyclodextrin glycosyltransferases (Klein & Schulz, 1991), accessory modules of certain cellulases (Juy *et al.*, 1992), proteins homologous to the first domain in growth hormone receptor (DeVos *et al.*, 1992) and probably interferon receptors (Bazan, 1990). None of these families bears sequence resemblance to each other, but in all cases conserved hydrophobicity patterns clearly indicate the presence of an antiparallel β-sheet. No other topology than that originally described in immunoglobulins has yet been reported for adhesive extracellular domains of this length. Taking all these facts together, an immunoglobulin-like fold for CUB is certainly a reasonable working hypothesis.

Given a common structural framework for all CUBs, there is still no indication for functional relations. Nevertheless, there are several indications of an involvement of all domains in developmental processes such as embryogenesis or organogenesis. All but one of the proteins containing both CUBs and a zinc-metalloprotease (UVS2, tolloid, Span and Bp10; Fig. 2) have been reported to be important in pattern formation during early embryogenesis (Sato & Sargent, 1989; Shimell *et al.*, 1991; Reynolds *et al.*, 1992; Lepage *et al.*, 1992). Interestingly, the only exception (Bmpl; Wozney *et al.*, 1988) was isolated from an extract that was capable of ectopic bone and cartilage formation (organogenesis), thus showing the connection between both processes at the molecular level. Two further undoubtedly developmentally regulated proteins are the A5 protein, which appears to be critical for targeting growing axons during nerve innervation (Takagi *et al.*, 1991), and Uegf (fibropellin), which is localized in the hyaline layer and mediates cell movements during the initial phase of gastrulation (Delgadillo-Reynoso *et al.*, 1989). The spermadhesins also fit in this classification because of their important role in fertilization (Sanz *et al.*, 1992*a*,*b*). The characterization of Asfp in seminal secretory vesicles (Wempe *et al.*, 1992) and its high sequence similarity to the spermadhesins point to a similar role for Asfp. For the TNF-inducible Tsg6 there are no direct biochemical data available that support a direct involvement in development, but

Figure 1. Multiple alignment of 31 CUB domains. The alignment was deduced using the program Clustal V (Higgens *et al.*, 1992) and it was afterwards refined "by hand". The protein names correspond to SWISSPROT codes (if already annotated). The beginning of a domain within the respective proteins is given in the 2nd column. Some variable regions are only indicated by the number of inserted amino acids. The β-strands are positioned according to hydrophobicity patterns (bold) or secondary structure predictions (see Bazan, 1990). A consensus line shows conserved features, for which up to 5 deviations per position have been tolerated. Capitals, strictly conserved amino acids; a, aromatic; h, hydrophobic; t, turn-like or polar ±, charged. The probable disulfide bonds between the cysteine residues are indicated in the bottom row.

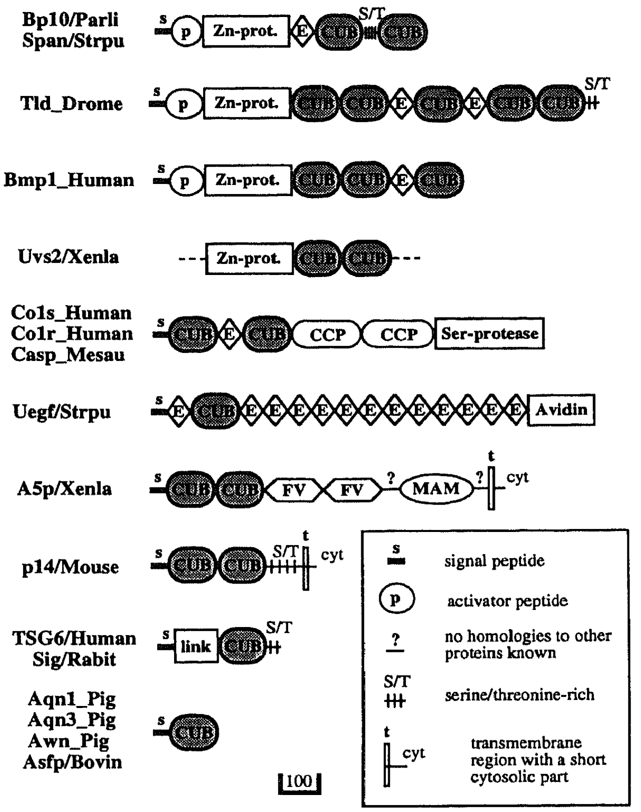

Figure 2. Modular architecture of the proteins containing CUB domains. Additional modules: Zn-prot, active zinc-metalloproteases; E, epidermal growth factor-like module; CCP, widespread repeat that is predominantly found in complement control proteins; Ser-protease, belongs to the serine protease family of the trypsin type, which is known to be shuffled together with other modules in various proteins of coagulation and complement; Avidin, module with similarity to the respective biotin-binding protein; FV, repeat that is also found in coagulation factors V and VIII as well as in milk fat globule protein; MAM, domain common to the receptors A5, meprin and tyrosine phosphatase μ (Beckmann & Bork, 1993); link, module in several proteoglucans and link protein.

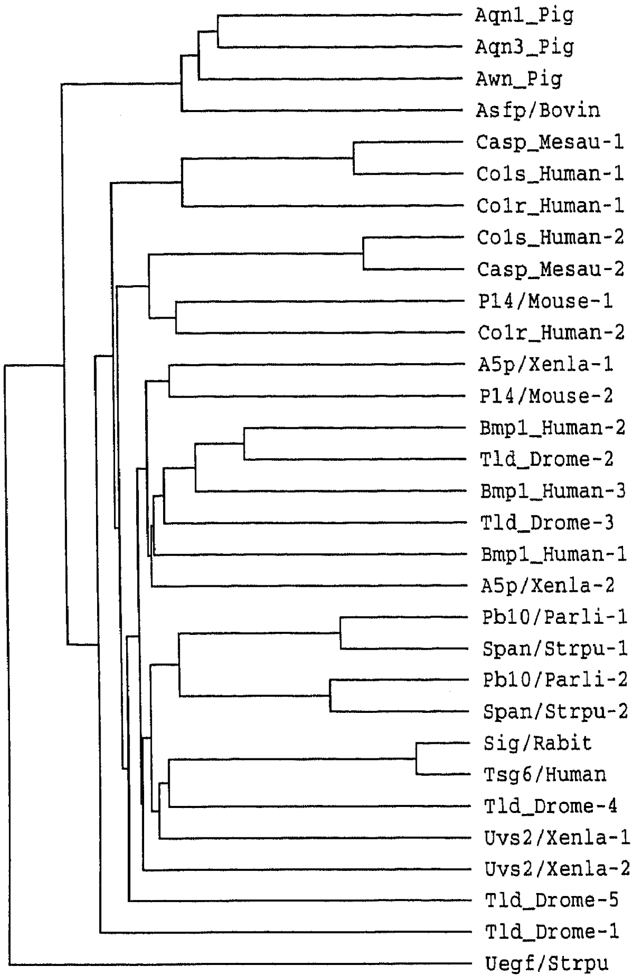

Figure 3. Dendrogram of all detected CUBs as produced by PILEUP of the GCG package (Devereux *et al.*, 1984). The resulting order of the sequences is also used in Fig. 1. The distance of Uegf is due to the missing C-terminal half of this domain, which remains to be sequenced. The spermadhesins and Asfp clearly form a subgroup as expected by their single domain architecture (Fig. 2). Interestingly, some of the multiple copies (e.g. in Span/Bp10 or Bmp1/tolloid) seem to have evolved by rather recent gene duplication.

TNF also induces the development of bactericidal granulomas, a process that, in the adult, mimics organogenesis (Kindler *et al.*, 1989). For p14 it has been reported that though the p14 mRNA was expressed at a high level in the astroglial cell line, from which it was isolated, there was no clear hybridization signal in astrocytes of the adult mouse (Lecain *et al.*, 1991). This hints at embryonic expression and awaits further exploration. We found no report of the involvement of C1s/C1r in developmentally regulated processes, but it is noteworthy that the very similar Casp, proposed as the hamster orthologue of human C1s (Bork, 1991), has been isolated from malignant embryo fibroblasts (Kinoshita *et al.*, 1989).

In summary, we define a family of modules for which we suggest analogous functions during developmental processes. At the molecular level this could be realized by specific carbohydrate binding as already shown for the spermadhesins (Sanz *et al.*, 1992*a*). Although the different members might have distinct binding specificities they probably share a structural framework (a β-barrel) similar to that of immunoglobulins. One test of this latter hypothesis would be the determination of the three-dimensional structure of distinct CUB domains. Indeed, NMR studies of the CUB domain in Tsg6 (I. D. Campbell, personal communication) as well as of Awn (E. Töpfer-Petersen, personal communication), a member of the most distant subgroup (Fig. 3), have already been initiated.

We thank I. D. Campbell and E. Töpfer-Petersen for discussions, as well as Chris Sander and Rebecca Wade for critical reading and helpful comments. G.B. was supported by Deutsche Forschungsgemeinschaft.

References

Bairoch, A. (1992). Prosite: a dictionary of protein patterns and sites. *Nucl. Acids Res.* **11**, 2013–2018.

Bazan, J. F. (1990). Structural design and molecular evolution of a cytokine receptor superfamily. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 6934–6938.

Beckmann, G. & Bork, P. (1993). A adhesive domain detected in several receptors. *Trends Biochem. Sci.* 40–41.

Bork, P. (1991). Complement components Clr/Cls, bone morphogenetic protein 1 and *Xenopus laevis* develepmentally regulated protein UVS.22 share common repeats. *FEBS Letters*, **282**, 9–12.

Bork, P. (1992). Mobile modules and motifs. *Curr. Opin. Struct. Biol.* **2**, 413–421.

Bork, P. & Doolittle, R. F. (1992). Proposed aquisition of an animal protein domain by bacteria. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 8990–8994.

Bork, P. & Grunwald, C. (1990). Recognition of different nucleotide-binding sites using a property pattern approach. *Eur. J. Biochem.* **191**, 347–358.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. & Sonnhammer, E. (1992). What's in a genome? *Nature (London)*, **358**, 287.

Delgadillo-Reynoso, M. G., Rollo, D. R., Hursh, D. A. & Raff, R. A. (1989). Structural analysis of the uEGF gene in the sea urchin *Strongylocentrotus purpuratus* reveals more similarity to vertebrate than to invertebrate genes with EGF-like repeats. *J. Mol. Evol.* **29**, 314–327.

Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucl. Acids Res.* **12**, 387–395.

DeVos, A. M., Ultsch, M. & Kossiakoff, A. A. (1992). Human growth hormone and its receptor: Crystal structure of the complex. *Science*, **255**, 306–312.

Gagnon, J. & Arlaud, G. J. (1985). Primary structure of the A-chain of human complement classical pathway enzyme Clr. *Biochem. J.* **225**, 135–142.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 4355–4358.

Higgens, D., Bleasby, A. J. & Fuchs, R. (1992). Clustal V: improved software for multiple sequence alignment. *CABIOS*, **8**, 189–191.

Holmgren, A., Kuehn, M. J., Braenden, C.-I. & Hultgren, S. J. (1992). Conserved immunoglobulin-like features in a family of periplasmic pilus chaperones in bacteria. *EMBO J.* **11**, 1617–1622.

Jones, E. Y., Davis, S. J., Williams, A. F., Harlos, K. & Stuart, D. I. (1992). Crystal structure at 2·8 Å resolution of a soluble form of the cell adhesion molecule CD2. *Nature (London)*, **360**, 232–239.

Juy, M., Amit, A. G., Alzari, P. M., Poljak, R. J., Claeyssens, M., Beguin, P. & Aubert, J.-P. (1992). Three-dimensional structure of a thermostable bacterial cellulase. *Nature (London)*, **357**, 89–91.

Kindler, V., Sappino, A.-P., Grau, G. E., Piguet, P.-F. & Vassalli, P. (1989). The inducing role of tumor necrosis factor in the development of bactericidal granulomas during BCG infection. *Cell*, **56**, 731–740.

Kinoshita, H., Sakiyama, H., Tokunaga, K., Imajoh-Ohmi, S., Hamada, Y., Isono, K. & Sakiyama, S. (1989). Complete primary structure of a calcium-dependent serine proteinase capable of degrading extracellular matrix proteins. *FEBS Letters*, **250**, 411–415.

Klein, C. & Schulz, G. E. (1991). Structure of cyclodextrin glycosyltransferase refined at 2·0 Å solution. *J. Mol. Biol.* **217**, 737–750.

Leahy, D. J., Hendrickson, W. A., Aukhil, I. & Erickson, H. P. (1992). Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science*, **258**, 987–991.

Lecain, E., Zelinika, D., Laine, M.-C., Rhyner, T. & Pessac, B. (1991). Isolation of a novel cDNA corresponding to a transcript expressed in the Choroid Plexus and Leptomeninges. *J. Neurochem.* **56**, 2133–2138.

Lee, T. H., Wisniewski, H.-G. & Vilcek, J. (1992). A novel secretory tumor necrosis factor-inducible protein (Tsg6) is a member of the family of hyaluronate binding proteins, closely related to the adhesion receptor CD44. *J. Cell. Biol.* **116**, 545–557.

Lepage, T., Ghiglione, C. & Gache, C. (1992). Spatial and temporal expression pattern during sea urchin embryogenesis of a gene coding for a protease homologous to the human protein Bmp-1 and to the product of the *Drosophila* dorso-ventral patterning gene tolloid. *Development*, **114**, 147–164.

Leytus, S. P., Kurachi, K., Sakariassen, K. S. & Davie, E. W. (1986). Nucleotide sequence of the cDNA coding for human complement Clr. *Biochemistry*, **25**, 4855–4863.

Main, A. L., Harvey, T. S., Baron, M., Boyd, J. & Campbell, I. D. (1992). Solution structure of the fibronectin type III module. *Cell*, **71**, 671–678.

Pearson, W. R. & Lipmann, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Nat. Acad. Sci. U.S.A.* **85**, 3338–3342.

Reynolds, S., Angerer, L. M., Palis, J., Nasir, A. & Angerer, R. C. (1992). Early mRNAs, spatially restricted along the animal-vegital axis of sea urchin embryos, include one encoding a protein related to tolloid and Bmp-1. *Development*, **114**, 769–786.

Rost, B. & Sander, C. (1992). Jury returns on structure prediction. *Nature (London)*, **360**, 540.

Ryu, S.-E., Kwong, P. D., Truneh, A., Porter, T. G., Arthos, J., Rosenberg, M., Dai, X., Xuong, N., Axel, R., Sweet, R. W. & Hendrickson, W. A. (1990) Crystal structure of an HIV-binding recombinant fragment of human CD4. *Nature (London)*, **348**, 419–426.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Sanz, L., Calvete, J. J., Mann, K., Schaefer, W., Schmid, E. R. & Töpfer-Petersen, E. (1991). The amino acid sequence of Aqn-3, a carbohydrate-binding protein isolated from boar sperm. *FEBS Letters*, **291**, 33–36.

Sanz, L., Calvete, J. J., Mann, K., Schaefer, W., Schmid, E. R. & Töpfer-Petersen, E. (1992*a*). The complete primary structure of the boar spermadhesin Aqn-1, a carbohydrate-binding protein involved in fertilization. *Eur. J. Biochem.* **205**, 645–652.

Sanz, L., Calvete, J. J., Mann, K., Schaefer, W., Schmid, E. R., Amselsgruber, W., Sinowatz, F., Ehrhard, M. & Toepfer-Petersen, E. (1992*b*). The complete primary structure of spermadhesin Awn, a zona pellucida-binding protein isolated from boar spermatozoa. *FEBS Letters*, **300**, 213–218.

Sato, S. M. & Sargent, T. D. (1989). Molecular approach to dorsoanterior development in *Xenopus laevis*. *Develop. Biol.* **137**, 135–141.

Shimell, M. J., Ferguson, E. L., Childs, S. R. & O'Connor, M. B. (1991). The *Drosophila* dorso-ventral patterning gene tolloid is related to human bone morphogenetic protein 1. *Cell*, **67**, 469–481.

Takagi, S., Hirata, T., Agata, K., Mochii, M., Eguchi, G. & Fujisawa, H. (1991). The A5 antigen, a candidate for the neuronal recognition molecule, has homologies to complement components and coagulation factors. *Neuron*, **7**, 295–307.

Tosi, M., Duponchel, C., Meo, T. & Julier, C. (1987). Complete cDNA sequence of human complement C1s and close physical linkage of the homologous genes C1s and C1r. *Biochemistry*, **26**, 8516–8524.

Wempe, F., Einspanier, R. & Scheit, K. H. (1992). Characterization by cDNA cloning of the mRNA of a new growth factor from bovine seminal plasma: acidic seminal fluid protein. *Biochem. Biophys. Res. Commun.* **183**, 232–237.

Wozney, J. M., Rosen, V., Celeste, A. J., Mitsock, L. M., Whitters., M. J., Kriz, R. W., Hewick, R. M. & Wang, E. A. (1988). Novel regulators of bone formation: Molecular clones and activities. *Science*, **242**, 1528–1534.

Edited by F. Cohen