Volume 327, number 2, 125–130 FEBS 12753 July 1993

Minireview

The modular architecture of a new family of growth regulators related to connective tissue growth factor

Peer Bork*

Max-Delbrück-Centre for Molecular Medicine, D-13125 Berlin-Buch, Germany

Received 18 May 1993; revised version received 8 June 1993

Recently, several groups have characterized and sequenced members of a new family of growth regulators (originally called *cef10*, connective tissue growth factor, *fisp-12*, *cyr61*, or, alternatively, βIG-M1 and βIG-M2), all of which belong to immediate-early genes expressed after induction by growth factors or certain oncogenes. Sequence analysis of this family revealed the presence of four distinct modules. Each module has homologues in other extracellular mosaic proteins such as Von Willebrand factor, *slit*, thrombospondins, fibrillar collagens, IGF-binding proteins and mucins. Classification and analysis of these modules suggests the location of binding regions and, by analogy to better characterized modules in other proteins, sheds some light onto the structure of this new family.

Mosaic protein; Homology; Extracellular module; Sequence analysis; Growth factor

1. INTRODUCTION

Cells can express a number of genes within minutes as the result of stimulation by growth factors or transforming oncogenes [1]. The majority of these immediate-early genes are intranuclear DNA regulatory proteins as well as transcription factors [2]. A second group comprises secreted, extracellular proteins which are needed for coordination of complex biological processes such as differentiation and wound healing [3].

Recently, several novel highly related proteins have been characterized which belong to the latter group of immediate-early genes. The first protein of this family, *cef10* from chicken, has been detected after induction by the viral oncogene $pp60^{v\text{-}src}$ [4]. A close relative, mouse *cyr61*, is rapidly activated by serum or platelet derived growth factor (PDGF) [5]. The overall amino acid identity between *cef10* and *cyr61* is as high as 83%, suggesting that both are orthologous genes in mouse and chicken. A third member has been termed human connective tissue growth factor (CTGF [6]). It is the major mitoattractant secreted by human umbilical vein endothelial cells and competes with PDGF for a particular cell surface receptor [6]. A fourth immediate-early protein of this family, *fisp-12*, was also found to be induced by serum and has been shown to be expressed in several tissue types of adult mice [3]. Because of the high sequence similarity to human CTGF (94% amino acid identity) *fisp-12* is probably the mouse orthologue of human CTGF. Interestingly, two recently characterized mouse genes that are induced by cell stimulation with transforming growth factor β (TGF-β), called βIG-M1 and βIG-M2 [7], are identical to *cyr61* and *fisp-12*, respectively. A chicken gene, *nov*, normally arrested in adult kidney cells, but overexpressed in myeloblastosis-associated virus type 1-induced nephroblastomas, was found to be yet another member of this emerging family of growth regulators [8] (here called CCN family: for **C**TGF; **c***ef10*/*cyr61* and **n***ov*). A dendrogram (Fig. 1) suggests that *nov* is not the chicken orthologue of mouse and human CTGF. Indeed, first sequencing results of a likely human orthologue of *nov* and its mapping to chromosome 8 as well as localization of human CTGF on chromosome 6 [9] reveal the presence of at least three paralogues; interestingly *fisp-12* (the mouse orthologue of human CTGF) has been mapped to the [A3-B1] region of chromosome 10 [3].

The sequence length of the five different proteins varies between 348 and 379 amino acids including the signal sequences. 38 completely conserved cysteines are clustered in two segments (22 and 16, respectively) sep-

Correspondence address. P. Bork, EMBL, Meyerhofstr. 1, D-69117 Heidelberg, Germany.

Abbreviations: CTGF, connective tissue growth factor; CCN, family of growth regulators comprising *cef10*/*cyr61*, CTGF and *nov*; IBPs, insulin-like growth factor binding proteins; IGF, insulin-like growth factor; PDGF, platelet derived growth factor; VWC, Von Willebrand factor type C module; TSP1, thrombospondin type 1 repeat; CT, C-terminal module; TGF-β, transforming growth factor β.

**Present address*: EMBL, Meyerhofstr. 1, D-69117 Heidelberg, Germany.

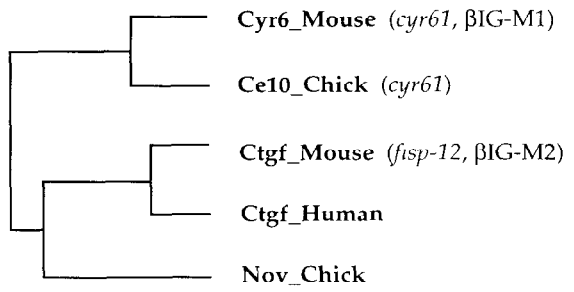

Fig. 1. Dendrogram of the members of the CCN family. The distances between the subclusters in accordance with the corresponding species supports the presence of (at least) three paralogues in vertebrates.

arated by a region which varies greatly in length and amino acid composition and which might be result of different splicing variants [6].

Because of the high level of sequence similarity among the members of the CCN family, they probably have most of their molecular functions in common. Thus, functional information provided for one member can be transferred to other proteins of the family. For example, *cyr61* is known to interact with both cell surfaces and the extracellular matrix, and it binds heparin with high affinity [10]. Similar binding activities can be anticipated for other family members, whereas specific interactions such as binding of CTGF to a defined PDGF receptor [6] might be a unique feature of this particular protein.

Apart from induction after stimulation by serum, growth factors or oncogenes, the normal expression pattern and organ specificity has been studied for several members. For example, in adult *nov* is found in large amounts in lung and brain, but in embryos it is only expressed in kidney [8]. *fisp-12* has an expression pattern similar to *cyr61* and is also most abundant in adult lung. However, this is still 10–20 times lower than in stimulated cells [3]. The expression of *cyr61* during embryogenesis has been studied more detailed and has been mainly assigned to the developing cartilaginous skeleton [11].

In this review the results of a comparative sequence analysis of the CCN family is presented. The presence of four distinct structural modules covering nearly the whole molecule shows that the CCN family members are genuine mosaic proteins (for review see [12] and refs. therein). Their modularity (Fig. 2) is used to suggest several functional sites.

2. MODULE 1 – AN INSULIN-LIKE GROWTH FACTOR BINDING DOMAIN?

Database searches with *cyr61* have revealed a strong local similarity to members of the low weight insulin-like growth factor-binding proteins (IBPs [5]). This local motif (GCGCCxxC) is well-conserved in most of the IBPs (Fig. 3) and is thought to be involved in IGF binding [13]. Low weight IBPs are secreted proteins of 24–30 kDa which form complexes with other extracellular proteins including IGF itself [13]. In addition to the region around the similaity to the CCN family they contain another domain (Fig. 2) related to repeats in thyroglobulin, nidogen, gastrointesinal tumor-associated antigens and the γ-chains of the H2 class II histocompatibility complex [13]. Thus, IBPs appear to be modular proteins and it can be expected that the N-terminal cysteine-rich region has homologues in other extracellular proteins. Indeed, the local region similar

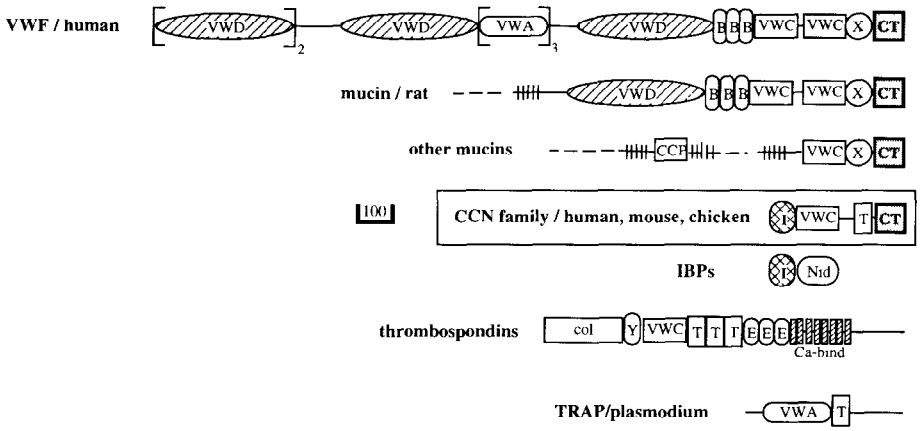

Fig. 2. Some relations of the modules identified in the CCN family. VWA–D,B, Von Willebrand factor type A–D repeats; col, module similar to some globular domains in collagens; T, thrombospondin type 1 repeat (TSP1); CT, a new C-terminal module in several matrix proteins; E, epidermal growth factor-like; I, IGF binding module; CCP, module frequently found in complement control proteins; Nid, module found in nidogen and other proteins; +++, small repeats; X,Y, vertical lines, not classified yet A unit of 100 amino acids is given for length comparison. All members of the CCN family have been subjected to a number of tools (see [39] for the searching scheme). The significance of the presented homologies has been assigned by using different heuristic methods [40]. For all modules FASTA and TFASTA searches [15] have been carried out and the results have been verified by pattern matching [41] and profilesearch methods ([42]; Thompson, J., Higgins, D. and Gibson, T., unpublished). Three of the four modules in the CCN family have been localized by screening our pattern library for modules of extracellular proteins [21]

Module 1 - IBP-like domain

```
sec.struct.                                    bbbb                                                    bbbb
Ce10_Chick  24 SPCPAVC......QCPAA-( 3)-CAPG..VGLVPDGCGCCKVCAKQLNEDCSRT.QPCDHTKGLECNFGASPAATN.......GICRA
Cyr6_Mouse  24 STCPAAC......HCPLE-( 3)-CAPG..VGLVRDGCGCCKVCAKQLNEDCSKT.QPCDHTKGLECNFGASSTALK.......GICRA
Ctgf_Mouse  26 QDCSAQC......QCAAE-( 4)-CPAG..VSLVLDGCGCCRVCAKQLGELCTER.DPCDPHKGLFCDFGSPANRKI.......GVCTA
Ctgf_Human  27 QNCSGPC......RCPDE-( 4)-CPAG..VSLVLDGCGCCRVCAKQLGELCTER.DPCDPHKGLFCDFGSPANRKI.......GVCTA
 Nov_Chick  29 AACPRPCG....GRCPAE-( 3)-CAPG..VPAVLDGCGCCLVCARQRGESCSPL.LPCDESGGLYCDRGPEDGGGA..... .GICMV
Ibp6_Human  25 VHCE.PCDEKALSMCPPS-( 3)-C.....ELVKEPGCGCCMTCALAEGQSCGVYTERC..AQGLRCLPRQDEEKPLHALLHGRGVCLN
Ibp3_Human  38 VRCE.PCDARALAQCAPP-( 3)-CA....ELVREPGCGCCLTCALSEGQPCGIYTERC..GSGLRCQPSPDEARPLQALLDGRGLCVN
Ibp1_Human  28 WQCA.PCSAEKLALCPPV-( 3)-CS....EVTRSAGCGCCPMCALPLGAACGVATARC..ARGLSCRALPGEQQPLHALTRGQGACVQ
Ibp4_Human  25 IHCP.PCSEEKLARCRPP-( 2)-CE....ELVREPGCGCCATCALGLGMPCGVYTPRC..GSGLRCYPPRGVEKPLHTLMHGQGVCME
Ibp2_Human  43 FRCP.PCTPERLAACGPP-(18)-CA....ELVREPGCGCCSVCARLEGEACGVYTPRC..GQGLRCYPHPGSSELPLQALVMGEGTCEK
Ibp5_Human  27 ARCP.GCGQGVQAGCPGG-( 0)-CVEEEDGGSPAEGCAEAEGCLRREGQECGVYTPNC..APGLQCHPPKDDEAPLRALLLGRGRCLP
consens          Ct  C       Cttt      C       h  tGCGCC hCAht G  Ct h ttC  ttGL C  ttttt          GhC t
```

Fig. 3. Alignment of the IGF binding module with all known human IBP isoforms. The sequences were taken from public databases (SWISS-PROT, PIR, EMBL). The first row corresponds to SWISS-PROT codes if these available. Conserved features (bold) are indicated in the consensus line (C=cysteines, t=turn-like or polar [E, D, Q, N, K, R, T, S, P, G, A], h=hydrophobic [I, L, V, W, Y, F, M, A, G], a=aromatic [Y, F, W], o=S/T). Predicted β-strands (rows of b) [38] are shown above the alignment. A large flexible region is only indicated by the number of inserted amino acids (in brackets).

to CCN family can be extended in a way that covers the whole cysteine-rich, N-terminal domain of the IBPs, i.e. 12 cysteines (Fig. 2, Fig. 3). The homology can be verified by various sequence database search methods, but for relatively small modules, pattern and profile search algorithms are advisible. The pairwise amino acid identity between IBPs and CCN members varies between 21% and 38% over a length of about 70 residues. More than 50% of all pairwise comparisons between any member of the CCN family and IBPs result in similarities higher than 30%, which is clearly above the threshold of structural homology [14].

The similarity to IBPs narrows a location of a putative IGF binding site. Because cysteines usually form disulfide bridges to stabilize the backbone in extracellular modules, rather than participating in binding interactions, conserved charged residues close to the characteristic motif are good candidates for binding interactions (Fig. 3).

The presence of even one module in the N-terminus of the CCN family implies a modular architecture. This is supported by the oncogenic potential of an intact, but truncated *nov* polypeptide lacking the N-terminal domain [8]. Indeed, additional modules can be defined adjacent to the putative IGF binding one described above (Fig. 2).

3. MODULE 2 – A COMPLEX FORMING DOMAIN?

When only parts of the proteins are used as probes in homology searches, standard methods like FASTA [15] reveal candidates for which the sequence relation to the probe can then be verified by profile and pattern searches. The second module identified is located next to the putative IGF-binding domain and covers all 10 remaining cysteines of the first cluster (Fig. 2, Fig. 4). This module has already been classified and named after the Von Willebrand factor type C repeat (VWC), having been found originally in two copies in the large, multifunctional protein Von Willebrand factor (see [16] and refs. therein). It is also present in several heavily

Module 2 - VWC domain

```
sec.struct.                bbbb  bbb        bbbbb           bbbb
  Cyr6_Mouse    98 RPCEYN.SRIYQNGESFQPN...CKHQCTCID.......GAVGCIPL.CPQELSLPNLGCPNPRLVKV.SGQCCEEWVCD
  Ce10_Chick    98 RPCEYN.SKIYQNGESFQPN...CKHQCTCID.......GAVGCIPL.CPQELSLPNLGCPSPRLVKV.PGQCCEEWVCD
  Ctgf_Mouse   100 APCVFG.GSVYRSGESFQSS...CKYQCTCLD.......GAVGCVPL.CSMDVRLPSPDCPFPRRVKL.PGKCCKEWVCD
  Ctgf_Human   101 APCIFG.GTVYRSGESFQSS...CKYQCTCLD.......GAVGCMPL.CSMDVRLPSPDCPFPRRVKL.PGKCCEEWVCD
   Nov_Chick   104 DNCVFD.GMIYRNGETFQPS...CKYQCTCRD.......GQIGCLPR.CNLGLLLPGPDCPFPRKIEV.PGECCEKWVCD
 Vwf_Human-X  2255 TQCIGEDGVQHQFLEAWVPDHQPCQI.CTCLS.......GRKVNCTTQPCPTAKAPTCGLCEVARLRQ.NADQCCPEYECV
Vwf_Human-C1  2429 KVCVHR.STIYPVGQFWEEG...CDV.CTCTDMEDAVMGLRVAQCSQKPCE.......DSCRSGFTYVLHEGECCG..RCL
Vwf_Human-C2  2580 EACMLN.GTVIGPGKTVMID..VCTT.CRCMVQVGVIS.GFKLECRKTTC........NPCPLGYKEENNTGECCG..RCL
  Ca13_Human    30 GGCSHL.GQSYADRDVWKPE..PCQI.CVCDS.......GSVLCDDIICDDQE....LDCPNPEIPF...GECCA..VCP
  Ca13_Chick     1 GGCTHL.GQEYADRDVWKPE..PCQI.CVCDS.......GSVLCDDIICDDQE....LDCPNPEIPL...GECCP..VCP
  Ca12/Xenla    36 GSCVQD.GQRYSDKDVWKPE..PCQI.CVCDT.......GTVLCDEIICEES.....KDCPNAEIPF...GECCP..ICP
  Ca12/Human    32 GSCVQD.GQRYNDKDVWKPE..PCRI.CVCDT.......GTVLCDDIICEDV.....KDCLSPEIPF...GECCP..ICP
  Ca11_Chick    31 GSCVQD.GLTYNDKDVWKPE..PCQI.CVCDS.......GNILCDEVICEDT.....SDCPNAEIPF...GECCP..ICP
  Ca11_Human    38 ITCVQN.GLRYHDRDVWKPE..PCRI.CVCDN.......GKVLCDDVICDET.....KNCPGAEVPE...GECCP..VCP
  Ca25_Human    39 IACTQN.GQMYLNRDIWKPA..PCQI.CVCDN.......GAILCDKIECQDV.....LDCADPVTPP...GECCP..VCS
  Ths2_Mouse   318 SACVQE.GRIFAENETWVVD..SCTT.CTCKK.......FKTVCHQITCSP......ATCANPSFVE...GECCP..SCS
  Ths2_Chick   324 SVCWQD.GRVFADSESWIVD..SCTK.CTCQD.......SKIVCHQITCPPV.....LSCADPSFIE...GECCP..VCS
  Ths1/Human   316 PLCYHN.GVQYRNNEEWTVD..SCTE.CHCQN.......SVTICKKVSCPI......MPCSNATVPD...GECCP..RCW
consens             Ch t G  attt- a  t   Cth ChC t        t hh C th C          tCt t h     G-CCt  hC
```

Fig. 4. Alignment of the VWC module with corresponding regions in collagen chains α_1 (III), α_1 (II) and α_2 (V), thrombospondin 1 and 2, as well as Von Willebrand factor. Only some representatives of fibrillar collagens and thrombospondin isoforms have been included. For nomenclature see Fig. 3. Interestingly, a region which has been assigned to the type B modules in Von Willebrand factor [18] also matches the corresponding consensus patterns [41].

glycosylated mucins (see [17,18] and refs. therein), in thrombospondins (for recent review see [18]) and in the N-terminal propeptides of fibrillar collagens [19–22]. The VWC module has a length of about 70 residues. The pairwise sequence identities between the CCN family and corresponding modules in the other proteins are in no case less than 23% and can be as high as 41% amino acid identity over 70 residues (*nov*/N-terminus of mouse collagen α_1 (I) chain; Fig. 4). This degree of similarity is highly significant. Various binding functions have been assigned to thrombospondins and Von Willebrand factor, but the role of the VWC in the latter remains unclear. Apart from the family described here, all other proteins containing the VWC module are known to form oligomers. Furthermore, in Von Willebrand factor, the best characterized of these, the duplicated VWC module is thought to participate in oligomerisation. Curiously, it is not involved in the initial dimerisation step which requires a covalent link between the chains [23]. Assuming the presence of a dimerization domain (see below), the VWC modules of the CCN family might connect different chains to form larger complexes.

4. MODULE 3 – CELL ATTACHMENT VIA BINDING MOTIFS FOR SULFATED GLYCOCONJUGATES?

The C-terminal cysteine-rich part of the CCN family also contains two different modules, the first of which begins right after the variable region in the central portion of the molecule (Fig. 2). It contains a motif [7] first identified in thrombospondin (type I repeat) but later found in several other extracellular proteins including, properdin, circumsporozoite protein from several malaria proteins, TRAP, f-spondin, UNC5, antistatin and complement components of the membrane attachment complex (for recent reviews see [19,24]). Although the region around the conserved WSxCSxxCG motif (Fig. 5) appears to be variable, the similarity to the other proteins can be extended so that it includes 6 mostly conserved cysteines and covers about 60 residues (Fig. 5). The absence of specific cysteine pairs in some of the thrombospondin type I modules (correlated mutations) has been used to predict the location of disulfide bridges (see [24,25]; Fig. 5).

The motif is thought to be involved in binding to both soluble and matrix macromolecules [18], in particular to sulfated glycoconjugates (see [26] and refs. therein). Other experiments have shown the presence of a cell attachment site within the module with a direct participation of the conserved heparin-binding and sulfatide binding motifs [27–29].

5. MODULE 4 – A DIMERIZATION DOMAIN?

The remaining C-terminal domain (CT module) appears to be homologous to the C-termini in functionally and structurally different extracellular mosaic proteins [30,31]. The closest homologue seems to be *slit*, a protein involved in development of midline glia and commisural axon pathways in *Drosophila* [32]. More distant

Module 3 - TSP1 domain

```
sec.struct.                          bbbbbbb                bbbbb
CE10_CHICK  224 CIVQTTSWSQCSKTCGTGISTRVTN.........DNPDCK...LIKETRICEVRP.CG
CTGF_HUMAN  199 CLVQTTEWSACSKTCGMGISTRVTN ........DNASCR...LEKQSRLCMVRP.CE
CTGF_MOUSE  198 CLVQTTEWSACSKTCGMGISTRVTN.........DNTFCR...LEKQSRLCMVRP.CE
CYR6_MOUSE  227 CIVQTTSWSQCSKSCGTGISTRVTN.........DNPECR ..LVKETRICEVRP.CG
 NOV_CHICK  202 CIEQTTEWSACSKSCGMGFSTRVTN.........RNQQCE...MVKQTRLCMMRP.CE
TRAP_PLAFA  244 CGVW.DEWSPCSVTCGKGTRSRKRE.......ILHEGC.  ..TSEIQEQCEEER.CP
 CSP_PLAFA  339 KNSISTEWSPCSVTCGNGIQVRIKPGSA..NKPKDELDYE..NDIEKKICKMEK.CS
 CSP_PLAYO  296 SSQLTEEWSQCSVTCGSGVRVRKRKNV...NKQPENLTL...EDIDTEICKMDK.CS
 CO6_HUMAN   84 LGDF.GPWSDCD.PCI.EKQSKVRSVLR.PSQFGGQPCTE .PLVAFQPCIPSKLCK
 CO7_HUMAN   30 WDFY.APWSECN.GCT.KTQTRRRSVAV.YGQYGGQPCVG..NAFETQSCEPTRCCP
 C8A_HUMAN   41 LSNW.SEWTDCF.PCQ.DKKYRHRSLLQ.PNKFGGTICSG..DIWDQASCSSSTTCV
 C8B_HUMAN   67 LSSW.SSWTTCD.PCQ.KKRYRYAYLLQ.PSQFHGEPCNF..SDKEVEDCVTNRPCG
 CO9_HUNAN   24 MSPW.SEWSQCD.PCL.RQMFRSRSIEV FGQFNGKRCTD..AVGDRRQCVPTEPCE
 CO6_HUMAN   26 HYAW.TQWTSCSKTCNSGTQSRHRQIVV.DKYYQENFCEQICSKQETRECNWQR.CP
 CO6_HUMAN  562 WGCW.SSWSTCDATY...KRSRTRECNNPAPQRGGKRCEG..EKRQEEDC......
 CO7_HUMAN  503 WSCW.SSWSPCVQG....KKTRSRECNNPPPSGGGRSCVG..ETTESTQC.......
 C8A_HUMAN  541 WSCW.SSWSVCRAG....IQERRRECDNPAPQNGGASCPG..RKVQTQAC......
 C8B_HUMAN  548 WNCW.SNWSSCSGR....RKTRQRQCNNPPPQNGGSPCSG..PASETLDC.......
FSPO/RAT    504 MSEW.ITWSPCSVSCGMGMRSRERYVK...QFPDGSVCML..PTEETEKCTVNEECS
FSPO/RAT    561 VTEW.GEWDDCSATCGMGMKKRHRMVK..MSPADGSMCKA..ETSQAEKCMMPE CH
PROP_HUMAN   80 WSLW STWAPCSVTCSEGSQLRYRRCVGWNG.QCSGKVAPGTLEWQLQACEDQQCCP
PROP_HUMAN  139 WSGW.GPWEPCSVTCSKGTRTRRRACNHPAP.KCGGHCPG..QAQESEACDTQQVCP
PROP_HUMAN  259 WGPW.GPVSPCPVTCGLGQTMEQRTCNHPVPQHGGPFCAG. DATRTHICNTAVPCP
THBS_HUMAN  382 WSPW.SEWTSCSTSCGNGIQQRGRSCD....S.LNNRCEG..SSVQTRTCHIQE CD
THBS_HUMAN  438 WSHW.SPWSSCSVTCGDGVITRIRLCNSPSPQMNGKPCEG..EARETKACKKDA.CP
THBS_HUMAN  495 WGPW.SPWDICSVTCGGGVQKRSRLCNNPTPQFGGKDCVG..DVTENQICNKQD.CP
UNC5/CAEEL  277 WSSW.SDWSACSSSC...HRYRTRACTVPPPMNGGQPCFG .DDLMTQECPAQL.CT
consens          h  W ttWStCS tCt     R Rth       ttttC        tt C  tt Ct
S-S                     1   2          3           3           1     2
```

Fig. 5. Alignment of module 3 with selected TSP1 repeats in complement components C6, C7, C8a, C8b, and C9, circumsporozoite proteins and TRAP protein from *Plasmodium falciparum*, f-spondin, properdin, thrombospondin and UNC5 protein. The predicted disulfide bridge pattern [24,25] are indicated. For nomenclature see Fig. 3.

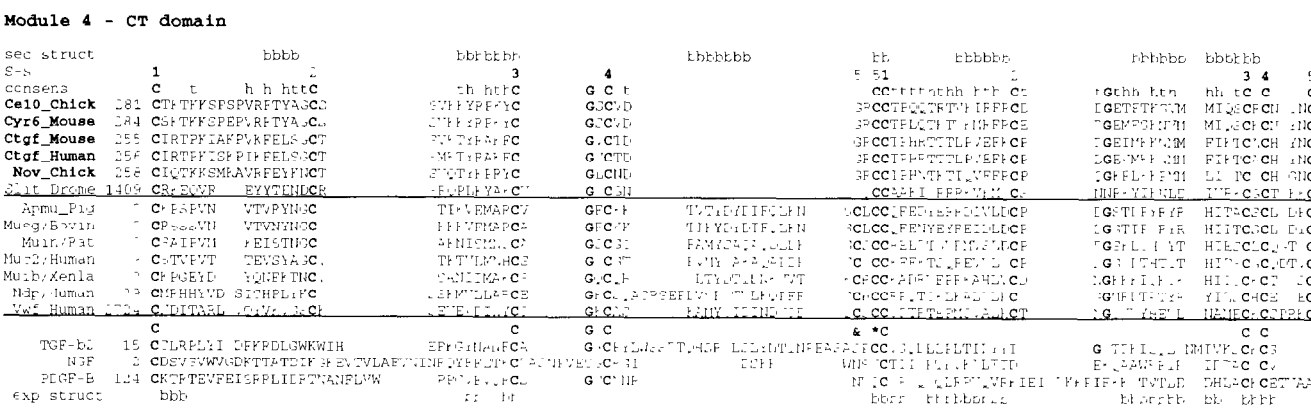

Fig. 6. Alignment of the CT domain with the C-termini of several extracellular mucins, Ndp (a candidate protein for Norrie disease), Von Willebrand factor and *slit*. For nomenclature see Fig. 3. For some proteins only fragments have been sequenced yet. The alignment of the CT domain is compared with the structure-based alignment [36] between tumor growth factor β_2, nerve growth factor and platelet derived growth factor (T. Meitinger et al., submitted). The β-strands common in the latter three growth factors are shown in the last line and are compared with the secondary structures predicted [38] for the CT-family (first line). Insertions/deletions are indicated by dots (within the CT module) or spaces (relative to TGF-β/NGF). The predicted disulfide bridges are marked. In the upper CT subgroup (growth regulators and *slit*) all cysteines form disulfide bridges and might dimerize similar to nerve growth factor [35,36]. The lower subgroup has an additional cysteine which, by analogy to TGF-β, might form an intermolecular disulfide bond. It could be either located in a position equivalent to TGF-β (*) or two residues before (&), in accordance with the fifth disulfide bridge of the upper subgroup.

related are Von Willebrand factor [17], several mucins [17,18] (see also module 2) which appear to be paralogues of each other (Fig. 2) and a protein involved in Norrie disease [31,33,34]. Two CT subgroups can be distinguished which are slightly different in their putative disulfide patterns (Fig. 6). According to this classification, only *slit* has a nearly identical spacing between the 10 cysteines of this module (Fig. 6); the similarity to the CCN family ranges between 23% and 26% over about 75 residues. Nevertheless, the other proteins containing this module can be significantly identified with profile and pattern searches. In all the proteins shown in Fig. 6, 8 out of the 10 cysteines are completely invariant; also numerous hydrophobic and turn-like positions remain conserved. Although the location of the CT module at the C-termini is striking, no function can yet be firmly assigned to this domain.

The spacing of six of the cysteines is similar (Fig. 6; T. Meitinger et al., submitted) to that in TGF-β which, in turn, belongs to a recently characterized family of structurally similar growth factors (for review see [35]). The latter form active dimers that bind to specific receptors. However, their mode of dimerization is completely different [35,36]. If the CT module has a fold similar to that of TGF-β, this would also explain the remaining disulfide bridges in structural terms (Fig. 6) and would even allow the prediction that the mucins, the protein responsible for Norrie disease and Von Willebrand factor dimerize via a particular disulfide bridge (Fig. 6). This hypothesis is in agreement with biochemical data, e.g. the last 151 residues of Von Willebrand factor (which include the CT module) have been shown to be involved in disulfide-linked dimerization [23]. Since most of the proteins containing the CT module are known to participate in complexes, this CT domain might indeed represent a dimerization domain and might even be involved in receptor binding.

6. CONCLUSION

In summary, the N-terminal putative IGF binding module of the CCN family is followed by a VWC domain; a variable region separates N-terminus and C-terminus which, in turn, also contains two modules, TSP1 and CT (Fig. 2). This modular architecture is strongly supported by the location of the known introns [3,37], 3 of which are exactly inserted between the 4 modules; the remaining one separates the signal peptide from the first module. Thus, the 38 conserved cysteines are spread over the four modules and are probably involved in disulfide bridges. The range of possibilities for the cysteine connections is narrowed, since they have to form disulfide bridges within the modules. For the TSP1 (Fig. 5) and the CT (Fig. 6) modules the disulfide pattern has already been deduced, either from correlated mutations of cysteines [24,25] or by homology (T. Meitinger et al., submitted). Since for each module a considerable number of homologous sequences is available, secondary structure predictions [38] allow a reasonable assignment of the β strands (Figs. 3–6). No α-helix has been predicted. Functional sites are probably located in loop regions in between the β strands. Interestingly, there are correlations between the presence of certain module types, e.g. the VWC module is often associated with the CT module in oligomers, but also frequently found together with the TSP1 repeat (Fig. 2).

Although equivalent structural modules in different

proteins might have distinct functions, the characterization of the modular architecture narrows the range of possibilities for the structure and function of the CCN family. A final prove for the functional suggestions, of course, can be only obtained by experiments.

Acknowledgements. I am grateful to T. Gibson and J. Thompson for introducing me in their new profile weighting method, to R F. Doolittle, R.C. Wade, V. Sorrentino and T. Meitinger for critical reading of the manuscript and to the Deutsche Forschungsgemeinschaft for financial support of the mosaic protein analysis.

REFERENCES

[1] Lau, L.F. and Nathans, D. (1991) in: Molecular Aspects of Cellular Regulation, Vol. 6 (Cohen, P. and Foulkes, J.G. eds.) Elsevier, Amsterdam, pp. 165–202.
[2] Bravo, R. (1990) Cell Growth Differ. 1, 305–309.
[3] Ryseck, R.-P., Macdonald-Bravo, H., Mattei, M.-G. and Bravo, R. (1991) Cell Growth Differ. 2, 235–233.
[4] Simmons, D.L., Levy, D.B., Yannoni, Y. and Erikson, R.L. (1989) Proc. Natl Acad. Sci USA 86, 1178–1182.
[5] O'Brien, T.P., Yang, G.P., Sanders, L. and Lau, L.F. (1990) Mol Cell. Biol. 10, 3569–3577.
[6] Bradham, D.M., Igarashi, A., Potter, R.L. and Grotendorst, G.R. (1991) J. Cell. Biol. 114, 1285–1294.
[7] Brunner, A., Chinn, J., Neubauer, M. and Purcho, A.F. (1991) DNA Cell. Biol. 4, 293–300.
[8] Joliot, V., Martinerie, C., Dambrine, G., Plassiart, G., Brisac, M., Crochet, J. and Perbal, B. (1992) Mol. Cell. Biol 12, 10–21.
[9] Martinerie, C., Viegas-Peuignot, E., Guenard, I., Dutrillaux, B., Nguyen, V.C., Bernheim, A. and Perbal, B. (1992) Oncogene 7, 2529–2534.
[10] Yang, G.P. and Lau, L.F. (1991) Cell Growth Differ. 2, 351–357.
[11] O'Brien, T.P. and Lau, L.F. (1992) Cell Growth Differ. 3, 645–654.
[12] Bork, P (1992) Curr. Opin. Struct. Biol. 2, 413–421.
[13] Kiefer, M.C., Masiarz, F.R., Bauer, D.M. and Zapf, J. (1991) J. Biol. Chem. 266, 9043–9049.
[14] Sander, C. and Schneider, R. (1991) Proteins 9, 56–68.
[15] Pearson, W.R. and Lipman, D. (1988) Proc. Natl. Acad. Sci USA 85, 2444–2448.
[16] Mancuso, D.J., Tuley, E.A., Westfield, L.A., Worrall, N.K., Shelton-Inloes, B.B., Sorace, J.M., Alevy, Y.G. and Sadler, J.E. (1989) J. Biol. Chem. 264, 19514–19527
[17] Xu, G., Huan, L.-J., Wang, D., Bennick, A., Fahim, R.E.F., Forstner, G.G. and Forstner, J F. (1992) J Biol. Chem. 267, 5401–5407.
[18] Probst, J.C., Hauser, F., Joba, W. and Hoffmann, W. (1992) *J Biol Chem.* 267, 6310–6316.
[19] Bornstein, P. (1992) FASEB J. 6, 3290–3299.
[20] Hunt, L.T. and Barker, W.C. (1987) Biochem. Biophys Res Commun. 144, 876–882
[21] Bork, P. (1991) FEBS Lett. 286, 47–54.
[22] Bork, P. (1992) FEBS Lett. 307, 49–54.
[23] Voorberg, J., Fontijn, R., Calafat, J., Janssen, H., Van Mourik, J A. and Pannekoek, H. (1991) J. Cell. Biol. 113, 195–205.
[24] Leung-Hagesteijn, C., Spence, A.M., Stern, B.D., Zhou, Y., Su, M.-W., Hedgecock, E M. and Culotti, J.G. (1992) Cell 71, 289–299.
[25] Smith, K.F., Nolan, K.F., Reid, K.B.M. and Perkins, S.J. (1991) Biochemistry 30, 8000–8008.
[26] Holt, G.D., Pangburn, M.K. and Ginsburg, V. (1990) J. Biol. Chem. 265, 2852–2855.
[27] Rich, K.A., George IV, F.W., Law, J.L and Martin, W.J. (1990) Science 249, 1574–1577
[28] Prater, C A., Plotkin, J., Jaye, D. and Frazier, W A. (1991) J Cell Biol. 112, 1031–1040.
[29] Guo, N.H., Krutsch, H.C., Negre, E., Vogel., T., Blake, D.A. and Roberts, D.D (1992) Proc. Natl. Acad. Sci. USA 89, 3040–3044.
[30] Rothberg, J.M. and Artavanis-Tsakonas, S. (1992) J Mol. Biol. 227, 367–370
[31] Meindl, A. et al. (1992) Nature Genet. 2, 139–143.
[32] Rothberg, J.M., Jacobs, J R., Goodman, C.S. and Artavanis-Tsakonas, S (1990) Gen. Dev. 4, 2169–2187.
[33] Berger, W. et al. (1992) Nature Genet. 1, 199–203.
[34] Chen, Z., Hendriks, R W., Jobling, M.A., Powell, J.F., Breakefield, X.O., Sims, K.B. and Craig, I.W. (1992) Nature Genet. 1, 204–208.
[35] McDonald, N.Q. and Hendrickson, W.A. (1993) Cell 73, 421–424.
[36] Swindells, M.B., Daopin, S., Cohen, G.H. and Davies, D (1992) Science 258, 1160–1162.
[37] Latinkic, B.V., O'Brien, T.P. and Lau, L.F. (1991) Nucleic Acids Res. 19, 3261 3267
[38] Rost, B. and Sander, C. (1992) Nature 360, 540.
[39] Bork, p., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. and Sonnhammer, E. (1992) Prot. Sci. 1, 1677–1690.
[40] Bork, P. & Sander, C. (1992) FEBS Lett. 300, 237–240.
[41] Rohde, K. and Bork, P. (1993). CABIOS 6, 183–189.
[42] Gribskov, M., McLachlan, A.D. and Eisenberg, D. (1987) Proc Natl. Acad. Sci. USA 84, 4355–4358.