Volume 300, number 3, 237–240 FEBS 10889 April 1992

A large domain common to sperm receptors (Zp2 and Zp3) and TGF-β type III receptor

Peer Bork and Chris Sander

EMBL, Meyerhofstr. 1, 6900 Heidelberg, Germany

Received 3 February 1992

A new family of mosaic proteins is defined by sequence analysis. The family is characterized by a 260 residue domain common to proteins of apparently diverse function and tissue specificity: sperm receptors Zp2 and Zp3, betaglycan (also called TGF-β type III receptor), uromodulin, as well as the major zymogen granule membrane protein (GP-2). The location of the common domain is similar with respect to putative transmembrane regions. The results lead to the hypothesis that this type of domain has a common tertiary structure and that there is a functional similarity in the recognition mechanism of the sperm receptor system and the TGF-β receptor complex.

Homology; Mosaic protein; Sperm receptor; TGF-receptor; Uromodulin

1. INTRODUCTION

Three different glycoproteins (ZP1, ZP2 and ZP3) of the zona pellucida, an extracellular matrix surrounding oocytes, are responsible for inducing the acrosome reaction (sperm exocytosis, for review see [1]). The detailed understanding of these molecular processes has a direct practical aspect in the context of development of safe contraceptive agents. Indeed, vaccination of a synthetic ZP3 peptide resulted in long term contraception in female mice [2]. Zp3 first binds to specific sperm proteins [3], thus mediating sperm contacts with the oocyte. Zp2 then acts as a second sperm receptor reinforcing the tight interactions [4]. Zp1 crosslinks Zp2 and Zp3, which both appear to form dimers, or perhaps higher oligomers (for review see [5]). Zp3 from different mammalia [6–8] as well as Zp2 from mouse [9] has been completely sequenced, but no sequence similarity between the two was observed.

We have now identified, by pattern-based sequence analysis, a large domain common to both sperm receptors. In addition, this domain can be found in several other receptor-like proteins such as TGF-β receptor III [10,11], uromodulin [12,13] and the major zymogen granule membrane glycoprotein GP-2 [14,15] defining a new family of mosaic proteins. Mosaic proteins have a modular architecture, as reviewed in [16,17]. The homologies presented here can be the basis for functional tests of the single modules in all proteins of the family. Conserved residues or segments provide useful hints for site-directed mutagenesis and DNA screening experiments.

Abbreviations: TGFR3, transforming growth factor β type III receptor; Zp2 (Zp3), sperm receptor Zp2 (Zp3); GP2, major zymogen granule membrane protein.

Correspondence address: P. Bork or C. Sander, EMBL, Postfach, Meyerhofstr. 1, 6900 Heidelberg, Germany.

2. MATERIALS AND METHODS

For homology searches SWISS-PROT [18] and PIR [19] sequence databases were used. Having detected a similarity between Zp2 and Zp3 using the program FASTA [20], the most conserved regions were described by property consensus patterns [21], which are sensitive and selective enough to detect remote relationship among extracellular mosaic proteins (e.g. [22]). As an additional test of the results every domain detected in this way was compared with *all* proteins of the sequence database using FASTA (k-tuple=1). Because of the apparent sequence flexibility and the high content of hydrophobic amino acids 1,000 hits were recorded and sorted in terms of the 'optimized' [20] scores. The multiple alignment was carried out using the program PILEUP of the GCG package [23]. A few additional gaps were introduced to align conserved cysteines.

When attempting to establish remote relationships, a significance estimate should be provided, especially if the proteins predicted to be homologous are not clearly related in terms of biological function. Evaluating statistically only pairwise alignments in a set of sequences, rather than the multiple align as a whole, can lead to erroneous underestimates of significance. Low-level resemblances become more significant as more members are added to the alignment [24]. Although there has been some progress in the mathematical treatment of significance in multiple sequence alignments, e.g. [25], insertions and deletions have yet to be included in the statistical estimates. We therefore assess our findings in three different empirical ways, in the context of three different homology search methods:

(1) FASTA. Even if no mathematical formula can be given, optimized scores [20] higher than 145 are thought by practioners to be significant, those higher than 100 to at least indicate possible homology, when comparing globular proteins with similar amino acid composition. Optimized scores take into account amino acid similarities as well as insertions/deletions and are more representative than pairwise identities alone. Table 1 shows both pairwise optimized scores and identities for all domains in this family.

(2) Rather than pairwise comparisons of sequences, the program PROFILESEARCH performs a multiple alignment driven by a profile, i.e. residue frequencies at each position [26]. The significance estimate in this method is provided by Z-scores (standard deviations above background), with values higher than 6.00 considered significant [26]. A database search with the derived profile yielded Z-scores higher than 11.00 for all of the predicted members of the domain family (data not shown). No other protein with a score higher than 5.20 was found.

(3) Whereas PROFILESEARCH takes into account the multiple sequence alignment over the entire length of the sequences, the program PAT [21] concentrates on the most conserved regions (motifs) only, representing them in terms of amino acid properties and allowing large gaps between the motifs ('consensus property patterns'). In this method, each of the 3 consensus regions independently appeared to be highly discriminating. At least 4 (6, 7 respectively) mismatches were recorded for other proteins not belonging to the family. A mismatch is a deviating amino acid property in any of the positions in the motif (for details see e.g. [27]).

Table 1

Pairwise similarities between the domains, in terms of optimized FASTA scores [20]

	urom	GP2D	GP2R	ZP2	ZP3M	ZP3H	TGFR3
urom	1,317	908	902	144*	54*	127	135*
GP2D		1,288	1,011	175	132*	133*	151*
GP2R			1,254	161	126*	149	142*
ZP2				1,413	158	171	103*
ZP3M					1,321	1,058	112
ZP3H						1,322	121*
TGFR3							1,620

The maximum possible score for a particular sequence is defined by the score for self-comparison along the diagonal.
* FASTA does not detect the full length of the domain due to insertions/deletions.

3. RESULTS AND DISCUSSION

Sequence analysis by FASTA revealed a homology between large segments of Zp2 and Zp3. Describing the most conserved regions of the common domain (marked in Fig. 1) by a consensus property pattern [21], further relationships to (i) transforming growth factor receptor III (TGFR3), (ii) uromodulin/Tamm-Horsfall protein, and (iii) the major granule membrane protein GP-2, became evident.

(i) TGFR3, also called betaglycan, is a component of the complex receptor system which mediates the numerous functions of transforming growth factor β (TGF-β). TGFR3 is thought to control the access of TGF-β to the signalling receptors [10,11]. (ii) Uromodulin was first discovered in the urine of pregnant women [28], but was later found to be a normal constituent of human urine, at lower concentrations. The protein fraction of this glycoprotein is identical to Tamm-Horsfall protein [12,13,29] which in turn is the major glycoprotein secreted by the mammalian kidney [30]. The functions of uromodulin/Tamm-Horsfall protein remain uncer-

```
Urom_human   332  EcGANDMKVSLGKcQLKSLGF-DKVFMYLSDSRcSGFNDRDNRDWVSVVTPARDGPcGTVLTRNETHATYSN  -( 6)-
GP2_dog      200  NcGAKEIQVSLDKcQLGGLGFGDEVIAYLRDWNcSNMMQREERNWISVTSPTQARAcGNILERNGTHAIYKN  -( 6)-
Gp2_rat      231  DcGANEIKVKLDKcLLGGLGFKEDIITYLNDRNcRGTMKDEPNNWVSTTSPVVANDcGNILENNGTQAIYRN  -( 6)-
Zp2_mouse    363  LcAQDGF-MDFEVYSHQTKPALNLDTLLVGNSScQPIFKVQSVGLARFHIPL--NGcGTRQKFEGDKVIYEN  -(10)-
ZP3_human     46  EcQEATLKVMVSKDLFGTGKLIRAADLTLGPEAcSPLVSMDTRDVVRFEVGL--HEcGNSMQVTDDALVYST  -(12)-
ZP3_mouse     45  EcLEAELVVTVSRDLFGTGKLVQPGDLTLGSEGcQPRVSVDT-DVVRFNAQL--HEcSGRVQMTKDALVYST  -(12)-
ZP3_hamster   45  EcLEAELVVTVSRDLFGTGKLIQPEDLTLGSENcRPLVSVAT-DVVRFKAQL--HEcSNRVQVTEDALVYST  -(12)-
TGFR3_rat    432  KcDHEKMVVAVDKDSFQTNGY-SGMELTLLDPScKAKMNGTH---FVLESPL--NGcGTRHRSTPDGVVYYN  -(36)-
                  C ppph h hp   hpp ph p  hhh p pCpp hp pp   h h ph ppCpp p ppp hhYpp

EIIIRDLNIKINFAcSYPLDMKVSLKTALQPMVSALNIRVGGTGMFTVRMALFQTPSYTQPYQGSSVTLST--EAFLYVGTMLDGGP SRFALLMTNcY
EFIIRDTILNINFQcAYPLDMKVSLQTALHPIVSSLNISVDGEGEFTVRMALFQDQSYISPYEGAAAVLAV--ESMLYVGAILEKGDTSRFNLLLRNcY
DFIIRDFLVNVNFQcAYPLDMNVSLQTALQPIVSSLNVDVGGAGEFTVTMALFQDQSYTHPYEGSKVLLPV--ENILYVGALLNRGDTSRFKLLLTNcY
NIVFRNSEFRMTVRcYYIRDSML-LN-AHVKGHPSPEAFVKP-GPLVLVLQTYPDQSYQRPYRKDEYPLVRYLRQPIYMEVKVLSRNDPNIKLVLDDcW
SIV-RTNRAEIPIEcRYPRQGNVSSQ-AILPTWLPFRTTVFSEEKLTFSLRLMEE-NWNAEKRSPTFHLC ---DAAHLQAEIHTGSHVPLRLFVDHcV
SIL-RTNRVEVPIEcRYPRQGNVSSH-PIQPTWVPFRATVSSEEKLAFSLRLMEE-NWNTEKSAPTFHLG----EVAHLQAEVQTGSHLPLQLFVDHcV
SIL-RTNRADVPIEcRYPRQGNVSSH-AIRPTWVPFSTTVSSEEKLVFSLRLMEE-NWNTEKLSPTSHLG----EVAYLQAEVQTGSHLPLLLFVDRcV
ETAPLSRAGVVVFNcSLR----------QLRNPSGFQGQLDGNATFNMELYNTDLFLVPSPGVFSVA-----ENEHVYVESVTKADQDLGFAIQTcF
phh  p  h h hpC h              ph  p ph hp  pphhh hphhpp  h  p  pp  hh     p  hh h h ppp  ph hhhppCh

ATPSS--NATDPLKYFII-QDRcPHT----RDSTIQVVENGESSQ---GRFSVQMFMFAGNY----DLVYLHcEVYLcDTMNEK------cKPTcSGT
ATPTK--DKTDPVKYFII-RNScPNQ----YDSTIHVEENGVSSE---SRFSVQMFMFAGNY----DLVFLHcEIHLcDSLNEQ------cQPCcSRS
ATPSG--DRNDIVKYFII-RNRcPNQ----RDSTINVEENGVSSE---SRFSVQMFRFAGNY----DLVFLHcEVYLcDSTTEQ------cQPScSTS
ATSSE--DPASAPQWQIV-MDGcEYE-LDNYRTTFHPAGSSAAHSCHYQRFDVKTFAFVSEARGLSSLIYFHcSALIcNQVSLDSPL---cSVTcPAS
ATPTP--DQNASPYHTIVDFHGcLVDGLTDASSAFKVPRPGPDT----LQFTVDVFHFANDSR---NMIYITcHLKVTLAEQDPDELNKAcSFSKPSN
ATPSPLPDPNSSPYHFIVDFHGcLVDGLSESPSAFQVPRPRPET----LQFTVDVFHFANSSR---NTLYITcHLKVAPANQIPDKLNKAcSFNKTSQ
ATPSP--LQTASPYHVIVDFHGcLVDGLSESFSAFQVPRPRPET----LQFTVDVFHFANSSP---NTIYITcHLKVTPANQTPDELNKAcSFNKSSK
LSPYS--NPDRMSDYTII-ENIcPKDSGVKFYSSKPVHFPIPHAEVDKKRFS---PLFKSVFN--TSLLFLHcELTLcSRKKGSLKL-PRcVTPFDAc
htt p  pppp   h In  p C  p       pphph  p  p    pPp   FhFhp       phhhhpCpph h  p p p       C hp ppp
```

Fig. 1. Multiple sequence alignment of the domain common to TGFR3, Zp2, Zp3, uromodulin and GP-2. The sequences were taken from SWISS-PROT [18] except for rat TGFR3 [10,11], dog GP-2 [15] and hamster Zp3 [7]. Amino acids conserved in all, except at most one, of the four subfamilies (uromodulin/GP2, Zp2, ZP3, TGFR3) are shown in bold face. A consensus line marks invariant amino acids (capitals) and conserved properties (h = hydrophobic, p = polar or turn-like and probably located at the protein surface). An intervening segment variable in length is only represented by the respective number of amino acids. The underlined segments were used for property pattern searches [21].

tain, but recent results support a role in preventing urinary tract infections by binding mannose-sensitive fimbriated microorganisms [31]. (iii) Glycoprotein GP- 2, the major component of pancreatic secretory granule membranes, is similar to uromodulin in sequence and domain structure [14], but lacks the three N-terminal EGF-like domains (Fig. 2).

The homologous region of about 260 residues common to these proteins (Figs. 1 and 2) contains eight strictly conserved cysteines, which may form disulfide bridges. The conservation of hydrophobicity, polarity or turn-forming tendency at numerous positions is consistent with a conserved three-dimensional structure (see consensus line in Fig. 1). In addition to the conserved cysteines, only a few aromatic or hydrophobic amino acids are absolutely invariant (Fig. 1), probably as a result of structural rather than functional constraints. Such a conservation pattern is typical of distantly related domains of mosaic proteins involved in binding functions [16,17].

The common domain occurs at a similar location relative to the putative membrane-spanning regions in each of these proteins (Fig. 2). This situation is suggestive of a possible common biological role of the domain. In this context, we note the following common biological properties of these proteins. (i) They all have been detected in soluble form, but (ii) also have features of integral membrane proteins in that they contain a long hydrophobic sequence segment at or near the C- terminus (Fig. 2). For three of them, membrane-bound forms, which are then further processed, have been characterized: uromodulin and GP-2 are known to contain a glycosylphosphatidylinositol membrane anchor [32,33], whereas the membrane-bound TGFR3 has a short cytoplasmic part homologous to that of endoglin [10,11,34]. In contrast, Zp2 and Zp3 have so far only been described as secreted proteins, but the short stretch of positively charged amino acids C-terminal to a hydrophobic region is typical of a small cytoplasmic extension of a membrane-bound form. Furthermore, all 5 proteins (iii) are heavily glycosylated, and (iv) appear in substantial amounts in the respective tissues (e.g. [10,14,32,35]).

The identification of a sizable common domain in these proteins is strongly suggestive of functional analogies. For the sperm receptors Zp2 and Zp3, previously not known to be structurally related, a common binding function to 95 kDa sperm proteins [3] is likely, with Zp3 binding first and Zp2 reinforcing the interaction. Our results also suggest a possible functional similarity in the recognition mechanism of the sperm receptor system and the TGF-β receptor complex.

REFERENCES

[1] Yanagimachi, R. (1988) in: The physiology of reproduction (Knobil, E. and Neill, J. eds.) Vol. 1, 135–185, Raven Press, New York.
[2] Miller, S.E., Chamow, S.M., Baur, A.W., Oliver, C., Robey, F. and Dean, J. (1989) Science 242, 935–938.
[3] Leytus, L. and Saling, P.M. (1989) Cell 57, 1123–1130.
[4] Bleil, J.D.C., Greve, J.M. and Wassarman, P.M. (1981) Dev. Biol. 86, 189–197.

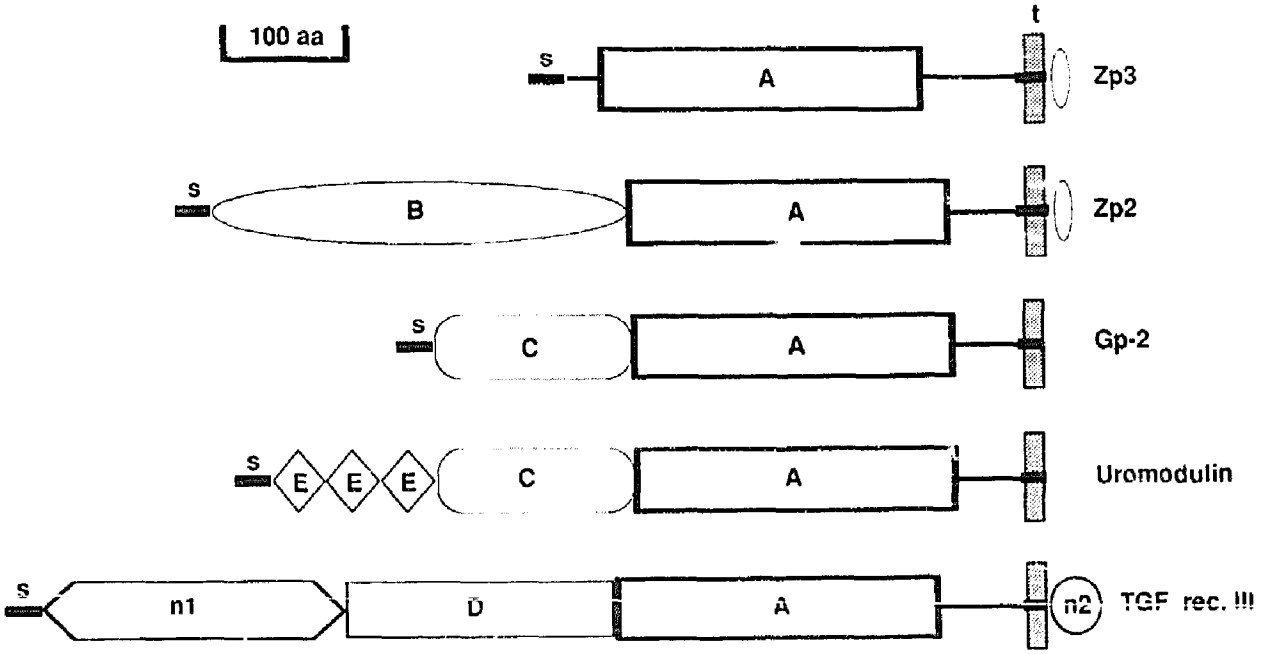

Fig. 2. Schematic representation of proteins containing the 260 residue domain. The domains differ in amino acid composition and degree of sequence conservation. A = domain, common to all four proteins (Fig. 1); B, C, D = segments for which so far no homology has been identified; E = epidermal growth factor (EGF-like) module; n1 (n2) = segments which are sequence-similar to endoglin [34] with 25% (63%) amino acid identity; s = signal sequence; t = hydrophobic segment, possibly membrane integrated.

[5] Wassarman, P.M. (1988) Annu. Rev. Biochem. 57, 415–442.
[6] Ringuette, M.J., Chamberlin, M.E., Baur, A.W., Sobieski, D.A. and Dean, J. (1988) Dev. Biol. 127, 287–295.
[7] Kinloch, R.A., Ruiz-Seiler, B. and Wassarman, P.M. (1990) Dev. Biol. 142, 414–421.
[8] Chamberlin, M.E. and Dean, J. (1990) Proc. Natl. Acad. Sci. USA 87, 6014–6018.
[9] Lian, L.-F., Chamov, S.M. and Dean, J. (1990) Mol. Cell. Biol. 10, 1507–1515.
[10] Lopez-Casillas, F., Cheifetz, S., Doody, J., Lane, W.S. and Massague, J. (1991) Cell 67, 785–795.
[11] Wang, X.F., Lin, H.Y., Ng-Eaton, E., Downward, J., Lodish, H.F. and Weinberg, R.A. (1991) Cell 67, 797–805.
[12] Pennicia, D., Kohr, W.J., Kuang, W.-J., Glaister, D., Aggarwal, B.B., Chen, E.Y. and Goeddel, D.V. (1987) Science 236, 83–88.
[13] Hession, C., Decker, J.M., Sherblom, A.P., Kumar, S., Yue, C.C., Mattaliano, R.J., Tizard, R., Kawashima, E., Schmeissner, U., Heletky, S., Chow, E.P., Burne, A.S. and Muchmore, A.V. (1987) Science 237, 1479–1484.
[14] Hoops, T.C. and Rindler, M.J. (1991) J. Biol. Chem. 266, 4257–4263.
[15] Fukuoka, S.-I., Freedman, S. and Scheele, G.A. (1991) Proc. Natl. Acad. Sci. USA 88, 2898–2902.
[16] Patthy, L. (1991) Curr. Opin. Struct. Biol. 1, 351–361.
[17] Bork, P. (1991) FEBS Lett. 286, 47–54.
[18] Bairoch, A. and Boeckmann, B. (1991) Nucleic Acids Res. 19, 2247–2249.
[19] Barker, W.C., George, D.G., Hunt, L.T. and Garavelli, J.S. (1991) Nucleic Acids Res. 19, 2231–2236.
[20] Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA 85, 3338–3342.
[21] Bork, P. and Grunwald, C. (1990) Eur. J. Biochem. 191, 347–358.
[22] Bork, P. (1991) FEBS Lett. 282, 9–12.
[23] Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. 12, 387–395.
[24] Doolittle, R.F. (1989) in: Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G.D. ed.) Plenum Publishing Corporation.
[25] Schuler, G.D., Altschul, S.F. and Lipman, D.J. (1991) Proteins 9, 180–190.
[26] Gribskov, M., McLachlan, A.D., Eisenberg, D. (1987) Proc. Natl. Acad. Sci. USA 84, 4355–4358.
[27] Bork, P. and Rohde, K. (1991) Biochem. J. 279, 908–910.
[28] Muchmore, A.V. and Decker, J.M. (1985) Science 229, 479–481.
[29] Tamm, I. and Horsfall, F. (1952) J. Exp. Med. 95, 71–97.
[30] Fabricius, T., Scott, D.M. and Kinne, R.K. (1989) Biol. Chem. Hoppe-Seiler 370, 151–158.
[31] Reinhart, H.H., Obedeanu, N., Robinson, R., Korzeniowski, O., Kaye, D. and Sobel, J.D. (1991) J. Urol. 146, 806–808.
[32] Rindler, M.J., Naik, S.S., Lin, N., Hoops, T.C. and Peraldi, M.-N. (1990) J. Biol. Chem. 265, 20784–20789.
[33] Rindler, M.J. and Hoops, T.C. (1990) Eur. J. Cell. Biol. 53, 154–163.
[34] Gougos, A. and Letarte, M. (1989) J. Biol. Chem. 265, 8361–8364.
[35] Saling, P.M. (1991) Biol. Reprod. 44, 246–251.