Protein Science (1993), *2*, 669–670. Cambridge University Press. Printed in the USA.

FOR THE RECORD

A trefoil domain in the major rabbit zona pellucida protein

PEER BORK
Max-Delbrück-Centre for Molecular Biology, 1115 Berlin-Buch, Germany

(RECEIVED December 28, 1992; REVISED MANUSCRIPT RECEIVED January 21, 1993)

Many extracellular proteins consist of various domains (modules) that often have defined binding functions (Bork, 1992). One of these modules is the so-called "trefoil" or P-domain (Thim, 1989; Tomasetto et al., 1990), which has a characteristic pattern of six conserved cysteines in a trefoil-like arrangement (Thim, 1989; Carr, 1992). This module has been described in several proteins (summarized in Fig. 1A) that have diverse biological activities, and it has been found in up to six copies per protein (Hauser & Hoffmann, 1992). Most of the proteins containing the trefoil domain are thought to be growth factors that also contribute to mucosal defense mechanisms, particularly after injury of the gastrointestinal tract (Suemori et al. [1991] and references therein). The function of the trefoil domain in this particular environment has been related to its resistance to proteolytic degradation (Suemori et al., 1991). Here I report a trefoil domain in a zona pellucida (ZP) protein indicating its widespread occurrence and suggesting a more general function for this module.

While updating our pattern library for extracellular modules (Bork, 1991), the consensus property pattern (Rohde & Bork, 1993) of known trefoil domains significantly matched a trefoil domain in the recently characterized major rabbit ZP protein (Schwoebel et al., 1991), whereas the random background of nonrelated proteins was separated by at least nine mismatches (Fig. 1; Rohde & Bork, 1993). All six cysteines and other consensus features are present in this protein (here called Zpx; see Fig. 1A). In addition, six of the known members have pairwise amino acid identities higher than 30% over about 50 residues, with human hsp protein and lysosomal α-glucosidase at the top (33% identity to each).

For two other regions of Zpx, Schwoebel et al. (1991) reported nonsignificant similarities to the Zp2 protein from mouse. I can add here that these two regions are both part of the recently characterized ZP module, which is also present in the zona pellucida proteins Zp2 and Zp3, TGF-β receptor type III (betaglycan), uromodulin, and the major zymogen granule membrane glycoprotein GP-2 (Bork & Sander, 1992). As for the trefoil domain, all conserved regions are present in the 310-residue-long domain of Zpx (data not shown). In addition, its closest relative (Zp2) has 37% amino acid identity within the ZP domain. The ZP module is located next to the trefoil domain, in agreement with the proposed modular architecture (Fig. 1B). No homology has been found for the N-terminal domain X (Fig. 1B), which is adjacent to the putative signal sequence (Schwoebel et al., 1991).

Together with the fact that the trefoil domain has also been identified in two intracellular enzymes—lysosomal α-glucosidase and sucrase-isomaltase (Tomasetto et al., 1990)—the presence of such a module in the heavily glycosylated zona pellucida indicates a more general role, such as specific binding to carbohydrates, and suggests a more widespread occurrence in various tissues.

References

Bork, P. (1991). Shuffled domains in extracellular proteins. *FEBS Lett. 286*, 47–54.

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

Bork, P. & Sander, C. (1992). A large domain common to sperm receptors (ZP2 and ZP3) and TGF-β type III receptor. *FEBS Lett. 300*, 237–240.

Carr, M.D. (1992). ^{1}H NMR-based determination of the secondary structure of porcine pancreatic spasmolytic polypeptide: One of a new family of "trefoil" motif containing cell growth factors. *Biochemistry 31*, 1998–2004.

Hauser, F. & Hoffmann, W. (1992). P-domains as shuffled cysteine-rich modules in integumentary mucin C.1 (FIM-C.1) from *Xenopus laevis*. *J. Biol. Chem. 267*, 24620–24624.

Rohde, K. & Bork, P. (1993). A fast, sensitive pattern matching algorithm for protein sequences. *CABIOS*, in press.

Schwoebel, E., Prasad, S., Timmons, T.M., Cook, R., Kimura, H., Niu, E.-M., Cheung, P., Skinner, S., Avery, S.E., Wilkins, B., & Dunbar, B.S. (1991). Isolation and characterization of a full-length

Reprint requests to Peer Bork at his present address: EMBL, Meyerhofstrasse 1, D-6900 Heidelberg, Germany.

```
A sec.struct.                                aaaaaaabbbb             bbbbb       mismatches
  Hsp/Human-1   30 CQC...SRLSPHNRTNCGF..PGITSDQCFDNGCCFDSSVTGV.......PWCFHPL      0    0
  Msp/Mouse-1   28 CRC...SRLTPHNRKNCGF..PGITSEQCFDLGCCFDSSVAGV.......PWCFHPL      0    0
  Spas_Pig-1    27 CRC...SRQDPKNRVNCGF..PGITSDQCFTSGCCFDSQVPGV.......PWCFKPL      0    0
  Itf/Rat       31 SQC....MAPTNVRVDCNY..PTVTSEQCNNRGCCFDSSIPNV.......PWCFKPL      0    0
  Xp4/Xenla-3  123 RDC...SAVEPKKRVNCGP..PGVSPDECIKNGCCFNSDVGGV.......PWCFKPE      0    0
  Xp4/Xenla-2   73 TIC...NPAEPKARVNCGY..PGITSQDCDKKGCCFNDTIPNV.......VWCYQPI      0    0
  Xp1/Xenla     30 EQC....SVERLARVNCGY..SGITPQECTKQGCCFDSTIQDA.......PWCFYPR      0    0
  Xp4/Xenla-4  173 LQC....AVLPKARINCGY..PDITMDQCYKKGCCYDSSESDS.......IWCFYPD      0    0
  Xp4/Xenla-1   25 YRC....GVKPKSRDNCGP..PGISPDECVKKGCCFDDSDPDS.......IWCYTPW      0    0
  Fimc/Xenla-4     GEC....KMEPSKRADCGY..PGITESQCRSKGCCFDSSIPQT.......KWCFYSL      0    0
  Fimc/Xenla-2     GEC....KMEPSKREDCGY..SGITESQCRTKGCCFDSSIPQT.......KWCFYTL      0    0
  Fimc/Xenla-5     ADC....KVAPSSRVDCGF..GGITADQCRQKNCCFDSSISGT.......KWCFYST      0    0
  Fimc/Xenla-3     ADC....KVEPSQRVDCGF..RGITADQCRQKNCCFDSSISGT.......KWCFYST      0    0
  Apeg_Xenla   347 EDC....KGDPFKRTDCGY..PGITEGQCKAKGCCFDSSIVGV.......KWCFFPA      0    0
  Spas_Xenla-1  21 QDC....SVAPNMRVNCGY..PTVTEADCRAVGCCFDSSILNT.......KWCFYNA      0    0
  Spas_Xenla-4 351 AEC....TVDPSVRTDCGY..PGITDKECREKGCCYDECIPDV.......IWCFEKA      0    0
  Spas_Xenla-2  72 LEC....SGDPTKRIDCGF..PRITEKQCILRGCCFDSSISGV.......KWCYART      0    0
  Spas_Xenla-3 303 PEC.......AADRVDCGY..SGITQADCEGKGCIFDSTIPET.......KWCFYTE      0    0
  Hsp/Human-2   80 DQC....VMEVSDRRNCGY..PGISPEECASRKCCFSNFIFEV.......PWCFFPN      0    0
  Msp/Mouse-2   78 EQC....VMEVSARKNCGY..PGISPEDCASRNCCFSNLIFEV.......PWCFFPQ      0    0
  Spas_Pig-2    77 EEC....VMQVSARKNCGY..PGISPEDCAARNCCFSDTIPEV.......PWCFFPM      0    0
  Ps2_Human     29 ETC....TVAPRERQNCGF..PGVTPSQCANKGCCFDDTVRGV.......PWCFYPN      0    0
  Fimc/Xenla-6     AMC....SGPPTKRRDCGY..PGISSSVCINRGCCWDNSVMNV.......PWCFYRT      0    0
  Fimc/Xenla-1     EHC....HVKPSKREMCGS..KGITKKQCKKKNCCFDPKGHGG.......IHCFHRK      0    0
  Zpx/Rabit    145 GLC...DSVPVQDRLPCAT..APISQEDCEELGCCHSSEEV.........NACYYGN      6    0
  Lyag_Human    80 TQC....DVPPNSRFDCAP.DKAITQEQCEARGCCYIPAKQGLQGAQMGQPWCFFPP      0    0
  Suis_Rat-1       xxxxxxxxxxxxxxxxCIP.EQSPTQAICEERGCCWRPWNNTV......IPWCFFAD     12   12
  Suis_Rabit-1  61 VSCPSELNEVVNERINCIP.EQSPTQAICAQRNCCWRPWNNSD......IPWCFFVD      0    0
  Suis_Human    61 GKCPNVLNDPVNVRINCIP.EQFPTEGICAQRGCCWRPWNDSL......IPWCFFVD      0    0
  Suis_Rat-2       FIVRWSQTFSDNEKFTCYPDVGTATEETDKQRGCLWQPVSGLS.....NVPPCYFPP      0    0
  Suis_Rabit-2 927 FRVQWDQTFLESEKITCYPDADIATQEKCTQRGCIWDTNTVNP.....RAPECYFPK      0    0
  consensus:         C          + tChh  tthot  C t tCCatt            haCaatt
  S-S bonds:         1             2           3    21                 3          1.   2.
```

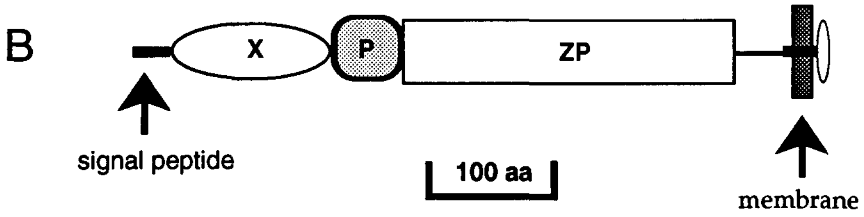

Fig. 1. A: Alignment of trefoil domains identified by our property pattern search method (Rohde & Bork, 1993). Apart from the fragment of Fimc (Hauser & Hoffmann, 1992), all proteins are already stored in the SWISSPROT (those with an underscore in the protein code) or PIR protein sequence databases. The beginning of the domains is given in the second column if the full sequence is available. The order of the sequences corresponds to a dendrogram (data not shown), beginning with the closest clusters. The secondary structure (a: α-helix, b: β-strands) determined by NMR (Carr, 1992) is shown above the alignment. Conserved features are indicated in the consensus line (C, conserved cysteines; t, turnlike 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). The putative disulfide bonds (Thim, 1989) are given by corresponding numbers. The lack of bond 1 in the second domain of sucrase-isomaltase (a fusion enzyme that arose from a gene duplication) and the low conservation of this domain between the rat and rabbit sequences suggest the loss of function. At the right, the number of mismatches is given for each sequence of the alignment. A mismatch is defined by the number of amino acid properties that deviate from the property consensus in a given position. This consensus is derived from a set of 11 physicochemical and steric properties. Combinations of those properties might be required or forbidden in each position (for details, see Rohde & Bork [1993]). The procedure includes an iterative step without (first run) and with (second run) inclusion of Zpx in the learning set (the multiple alignment). The relatively large number of mismatches for the first domain of rat sucrase-isomaltase (fragment) is because it lacks the first part of the domain. The best-scoring false positives have at least nine mismatches. **B:** The modular architecture of the 55-kDa zona pellucida protein is very similar to that of Zp2; only the domains X and P (trefoil) are substituted against a larger one in Zp2 for which no similarity to any other protein in sequence databases has yet been found. Because the ZP domain is followed by a membrane inserting element in all described proteins of this family (Bork & Sander, 1992), I predict, by analogy, a transmembrane region for one of the hydrophobic segments in Zpx (Schwoebel et al., 1991).

cDNA encoding the 55-kDa rabbit zona pellucida protein. *J. Biol. Chem. 266*, 7214-7219.

Suemori, S., Lynch-Devaney, K., & Podolski, D.K. (1991). Identification of rat intestinal trefoil factor: Tissue- and cell-specific member of the trefoil protein family. *Proc. Natl. Acad. Sci. USA 88*, 11017-11021.

Thim, L. (1989). A new family of growth factor-like peptides. *FEBS Lett. 250*, 85-90.

Tomasetto, C., Rio, M.-C., Gautier, C., Wolf, C., Hareuveni, M., Chambon, P., & Lathe, R. (1990). hSP, the domain-duplicated homolog of pS2 protein, is co-expressed with pS2 in stomach but not in breast carcinoma. *EMBO J. 9*, 407-414.