J. Mol. Biol. (1992) **227**, 367–370

Modularity of the Slit Protein

Characterization of a Conserved Carboxy-terminal Sequence in Secreted Proteins and a Motif Implicated in Extracellular Protein Interactions

Jonathan Marc Rothberg and Spyros Artavanis-Tsakonas

Howard Hughes Medical Institute
Departments of Cell Biology and Biology
Yale University
New Haven, CT 06511, U.S.A.

(Received 4 February 1992; accepted 28 May 1992)

Since our characterization of the *slit* cDNA sequence, encoding a protein secreted by glial cells and involved in the formation of axonal pathways in *Drosophila*, we have discovered that the protein contains two additional sequence motifs that are highly conserved in a variety of proteins. A search of the GenPept database with the 73 amino acids at the carboxy terminus of slit revealed that this region contains significant similarity to a carboxy-terminal domain found in six other exported proteins. This observation has allowed us to define a new carboxy-terminal protein motif. In addition, comparisons with a 202 amino acid domain residing between epidermal growth factor (EGF) repeats in slit shows this region to be conserved in laminin, agrin and perlecan and, strikingly, also to lie between EGF repeats in both agrin and perlecan. Our analysis suggests this motif is involved in mediating interactions among extracellular proteins. Consistent with our previous characterization of the slit protein, both new motifs are found only in extracellular proteins. The identification of these two conserved motifs in slit reveals that the entire 1469 amino acids of the protein are made up of modular regions similar to those conserved in other extracellular proteins.

Keywords: slit; agrin; laminin; protein sequence motifs; epidermal growth factor

Mutations in the *slit* locus result in the collapse of the regular scaffold of commissural and longitudinal nerve bundles in the *Drosophila* embryonic central nervous system (CNS†) (Rothberg *et al.*, 1988). An examination of the expression pattern of *slit* in the embryo demonstrates that the protein is expressed by a subset of specialized glial cells along the midline of the CNS. By immunoelectron microscopy, slit can be seen on the axons traversing these glial cells although it is absent from the cell bodies of these neurons, implying that slit is exported by the glia and distributed along these axons (Rothberg *et al.*, 1990). The embryonic localization, mutant phenotype, and homology of *slit* to both receptor-binding EGF-like ligands and adhesive glycoproteins, suggest that it may be involved in interactions between the midline glial cells, their extracellular environment, and the commissural axons that cross the midline.

The predicted protein product encoded by the *slit* locus contains four tandem arrays of a 24 amino acid leucine-rich repeat (LRR) with conserved flanking sequences surrounding these arrays (flank–LRR–flank), followed by two regions with EGF repeats (Fig. 1; Rothberg *et al.*, 1990). Each of these motifs has been shown to participate in protein–protein interactions as part of an extracellular domain in a variety of other proteins. As a consequence of alternative mRNA splicing, the locus encodes two distinct protein species of 1469 and 1480 amino acids, differing by 11 amino acids at the carboxy-terminal of the last EGF repeat. The existence of a putative signal sequence and the absence of a transmembrane domain suggest that slit is secreted, an observation supported by an analysis of its expression in tissue culture. In this report we further support and extend these observations by identifying a previously uncharacterized motif in

† Abbreviations used: CNS, central nervous system; EGF, epidermal growth factor; LRR, leucine-rich repeat; FIM-B.1, *Xenopus laevis* integumentary mucin; vWf, von Willebrand factor; ALPS, agrin, laminin, perlecan and slit motif.

0022-2836/92/180367-04 $08.00/0

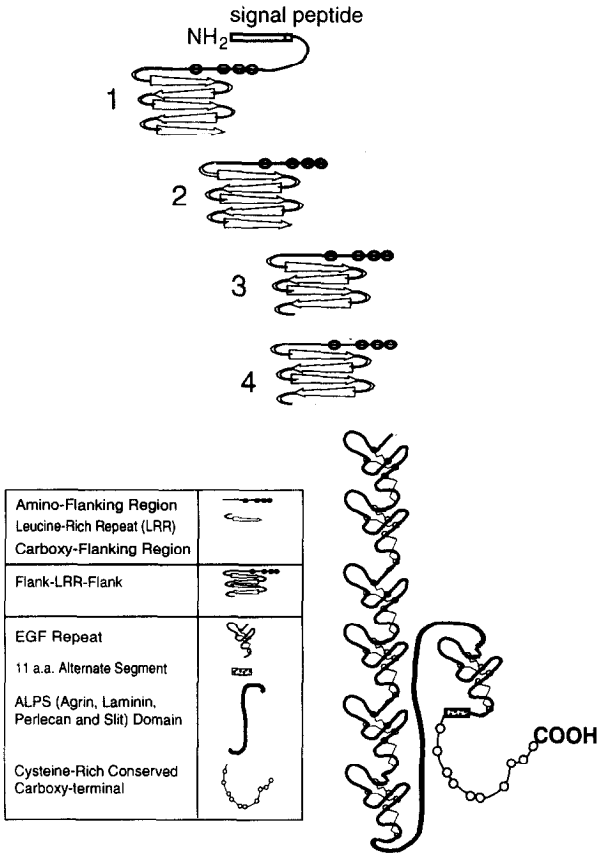

Figure 1. Schematic representation of the slit protein showing the arrangement of the 4 flank–LRR–flank structures, the 7 EGF repeats and their relationship to the conserved regions between the 6th and 7th EGF repeat, and the conserved carboxy-terminal motif (see key; adapted from Rothberg *et al.*, 1990). The putative signal sequence and amino- and carboxy-terminal ends of the protein are also indicated. The 4 consecutive flank–LRR–flank regions are composed of a tandem array of 4 or 5 24-amino acid leucine-rich repeats (LRRs) surrounded by conserved amino- and carboxy-flanking regions. These are followed by 6 EGF repeats separated from a 7th and final EGF repeat by 202 amino acids that are similar to regions in agrin, perlecan, and laminin. The 11 amino acid (a.a.) connecting segment, the result of differential splicing at the carboxy terminus of the 7th EGF repeat is shown. Following either the last EGF repeat or this segment, and located at the carboxy terminus of the protein, is a conserved carboxy-terminal cysteine-rich motif. As depicted here: LRRs are believed to form anti-parallel sheets in larger proteins; EGF motifs are modeled after the solution structure of human EGF; and tandem EGF-like repeats in other extracellular matrix proteins have been shown to be arranged in a rod-like conformation.

the carboxy-terminal region of slit that has been conserved in a corresponding position in otherwise unrelated exported proteins.

The conserved carboxy-terminal domain is characterized by eight invariant cysteine residues as well as other conserved amino acids (Fig. 2). Strikingly, in addition to this motif always residing at the extreme carboxy terminus, all the proteins containing it appear to be secreted. The proteins containing this conserved sequence motif, which otherwise share no significant sequence similarity with slit, can be subdivided into two groups. The first includes a family of proteins with overall sequence similarity to each other: CEF-10 (for chicken embryo fibroblast; Simmons *et al.*, 1989), *cyr61* (O'Brien, 1990) , CTGF (for connective tissue growth factor; Bradham *et al.*, 1991), *fisp-12* (Ryseck *et al.*, 1991) and *nov* (for nephroblastoma-overexpressed gene; Joliot *et al.*, 1992). The genes *cyr61* and *fisp-12* are likely to be the mouse homologs of chicken CEF-10 and human CTGF, respectively (O'Brien, 1990; and our observation). Interestingly, these four genes are inducible by either v-*src* or growth factors and, while secretion has only been demonstrated for CTGF and fisp-12, they all (including nov) contain a secretory signal sequence. The second group consists of a *Xenopus laevis* integumentary mucin (FIM-B.1) and human von Willebrand factor (vWF), which in addition to the aforementioned conserved carboxy-terminal region, contains a region of homology with antistasin (Probst, 1990). The vWF is synthesized and exported by platelet precursors and is involved in platelet to blood vessel adhesion while the FIM-B.1 protein is produced in the mucous glands where it is thought to be secreted and to function on the surface of the skin.

The presence of a conserved carboxy-terminal domain in otherwise unrelated secreted proteins raises the possibility that this region plays a specialized functional role either in the secretion process or in mediating specific associations with other proteins. It has been suggested that a regulatory sequence involved in the secretory routing of vWF may reside in a region that includes this conserved motif (Wagner *et al.*, 1991; Voorberg *et al.*, 1991). In addition, association with other extracellular proteins has been demonstrated or suggested for slit, vWF, and FIM-B.1.

With the characterization of the carboxy-terminal domain of slit as a conserved motif, the only region of the slit protein that has not been shown to be either associated with extracellular events or conserved in extracellular proteins is a stretch of approximately 200 amino acids separating the sixth and seventh EGF repeats (see Fig. 1). However, consistent with our characterization of slit as a modular extracellular protein, we have now discovered that this region contains a sequence motif that is conserved in other extracellular proteins (Fig. 3). Interestingly, this motif has also been implicated in protein–protein interactions. A comparison of this 202 amino acid sequence from slit with the GenPept database (Release 71.0; Pearson & Lipman, 1988) revealed a high degree of sequence similarity to the laminin A-chain as well as to regions of perlecan (Noonan *et al.*, 1991) and agrin (Rupp *et al.*, 1991). When this region of each of these proteins is aligned with the others they can be seen to have many invariant as well as conserved residues (Fig. 3). In recognition of the conservation of this motif in these proteins we will refer to it as

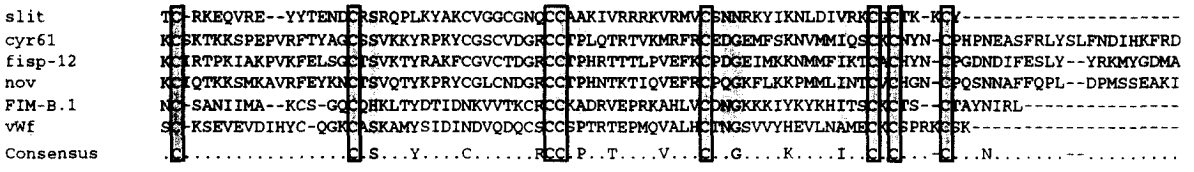

Figure 2. A conserved carboxy-terminal sequence motif in exported proteins. The amino acid sequence of a 73 amino acid long domain at the carboxy terminus of slit and five other secreted proteins is shown aligned above a consensus sequence. Invariant residues are boxed; residues at conserved positions are shaded. Eight cysteines constitute all the invariant positions. The consensus sequence is based on the presence of a specific amino acid residue at a given position greater than 50% of the time. The structure of the carboxy-terminal motif in slit is unknown, but the presence of conserved cysteine residues in these proteins suggests a common tertiary structure.

an ALPS (agrin, laminin, perlecan and slit) motif. It is remarkable that in perlecan and agrin this motif also lies between EGF repeats.

Laminin a three-chain basement membrane glycoprotein, is able to interact with both cell surface receptors and other members of the basement membrane. Interestingly, one of these components is perlecan, a large heparin sulfate proteoglycan believed to play an organizing role in the basement membrane. Perlecan, in addition to being able to interact with laminin, can self-aggregate. In laminin and perlecan the domains responsible for these interactions consist of repeats of the ALPS motif (Sasaki *et al.*, 1988; Noonan *et al.*, 1991). In laminin the motif is repeated contiguously five times forming the entire A-chain C-terminal globule. (This region is also conserved in merosin, a laminin A chain homolog; Ehrig *et al.*, 1990.) While, in perlecan, the motif is repeated three times with EGF-like repeats preceding and following the second iteration of the motif (Noonan *et al.*, 1991), this conservation is also observed in the human homolog of perlecan, referred to as the basement membrane heparin sulfate proteoglycan core protein (Kallunki Tryggvason, 1992). It is the second iteration of the motif that shows the greatest similarity to slit sequence (shown in Fig. 3).

Agrin, like slit and perlecan, contains the ALPS motif between EGF repeats. In agrin the conserved motif appears twice, lying between the first and

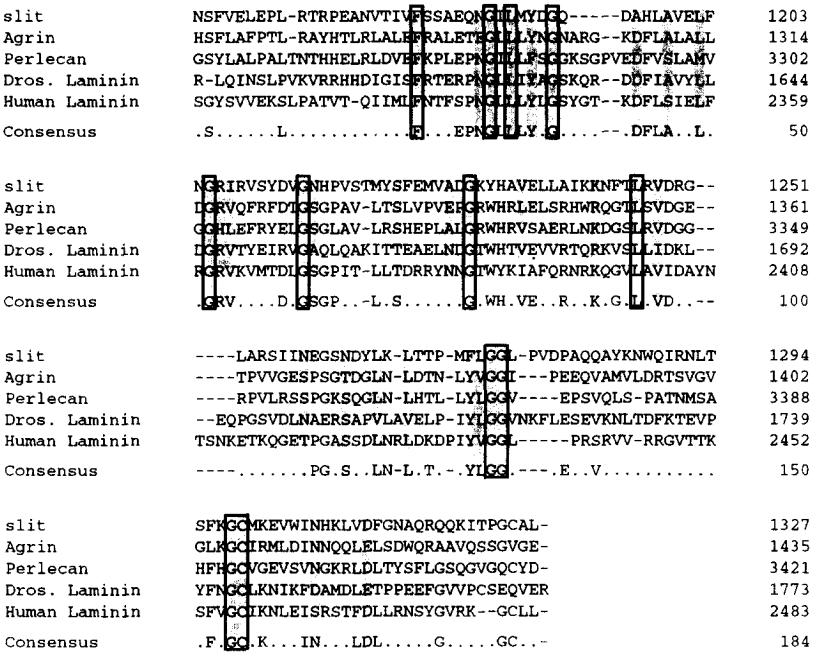

Figure 3. Conservation of sequence intervening between EGF repeats of slit and other extracellular matrix proteins. The amino acid sequences of rat agrin, mouse perlecan, and *Drosophila* and human laminin are shown below the corresponding slit sequence. Invariant residues are boxed; conserved positions are shaded. Numbers indicate the position of the amino acid sequences within their respective proteins. Note the predominance of glycine residues in the consensus sequence. In laminin the ALPS motif has been seen by electron microscopy to form a globular structure, and is believed to do the same in perlecan (Noonan *et al.*, 1991), suggesting that the invariant and conserved residues in the motif may facilitate the formation of a common structure. In both this and Fig. 2, the similarities to slit sequences were initially identified by searching version 71.0 of the GenPept database with the corresponding slit sub-sequences (FastA program; Pearson & Lipman, 1988). Initial FastA scores for the sequences identified were from 6 to 11 and from 7 to 17 standard deviations from the mean for the 2 searches, respectively. The multiple sequence alignments shown were derived using the Geneworks 2.0 program (IntelliGenetics, Inc., Mountain View, CA).

second, and third and fourth EGF repeats, respectively. The agrin protein, in addition to its ability to aggregate acetylcholine receptors on cultured muscle fibers, is also able to reorganize the surrounding extracellular matrix components (Nitkin & Rothschild, 1990). Interestingly, it is a region of the agrin protein containing this conservation, the carboxy terminal-most 94 kDa, that confers these functions. The slit protein is necessary for the development of glial cells that reside along the midline of the CNS and can be seen to decorate the axon pathways that traverse them (Rothberg *et al.*, 1990). The association of agrin with developing synapses and slit with growing axons raises the possibility that the function of the conserved ALPS motif in both proteins may be similar. In addition, the possibility that agrin's ability to organize the extracellular matrix is conserved in slit is raised by the observation that the slit protein becomes localized to points where extracellular matrix material linking muscles to the larval body wall will be deposited (Rothberg, 1991). These observations as well as its presence in laminin and perlecan suggest that, in addition to participating in protein interaction, this motif may play a general role in organizing the extracellular matrix.

References

Bradham, D. M., Igarashi, A., Potter, R. L. & Grotendorst, G. R. (1991). Connective tissue growth factor: a cysteine-rich mitogen secreted by human vascular endothelial cells is related to the SRC-induced immediate early gene product CEF-10. *J. Cell. Biol.* **114**, 1285–1294.

Ehrig, K., Leivo, I., Argraves, W. S., Ruoslahti, E. & Engvall, E. (1990). Merosin, a tissue-specific basement membrane protein, is a laminin-like protein. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 3264–3268.

Joliot, V., Martinerie, C., Dambrine, G., Plassiart, G., Brisac, M., Crochet, J. & Perbal, B. (1992). Proviral rearrangements and overexpression of a new cellular gene (nov) in myeloblastosis-associated virus type 1-induced nephroblastomas. *Mol. Cell. Biol.* **12**, 10–21.

Kallunki, P. & Tryggvason, K. (1992). Human basement membrane heparin sulfate proteoglycan core protein: A 467-kD protein containing multiple domains resembling elements of the low density lipoprotein receptor, laminin, neural cell adhesion molecules, and epidermal growth factor. *J. Cell. Biol.* **116**, 559–571.

Nitkin, R. M. & Rothschild, T. C. (1990). Agrin-induced reorganization of extracellular matrix components on cultured myotubes: Relationship to AChR aggregation. *J. Cell. Biol.* **111**, 1161–1170.

Noonan, D. M., Fulle, A., Valente, P., Cai, S., Horigan, E., Sasaki, M., Yamada, Y. & Hassell, J. R. (1991). The complete sequence of perlecan, a basement membrane heparan sulfate proteoglycan, reveals extensive similarity with laminin A chain, low density lipoprotein-receptor, and the neural cell adhesion molecule. *J. Biol. Chem.* **266**, 22939–22947.

O'Brien, T. P. (1990). Expression of cyr61, a growth factor-inducible immediate-early gene. *Mol. Cell. Biol.* **10**, 3569–3577.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Nat. Acad. Sci., U.S.A.* **85**, 2444–2448.

Probst, J. C. (1990). An integumentary mucin (FIM-B.1) from *Xenopus laevis* homologous to von Willebrand factor. *Biochemistry.* **29**, 6240–6244.

Rothberg, J. M. (1991) *slit*: an extracellular protein necessary for the development of midline glia and axon pathways of the central nervous system contains both EGF and flank-LRR-flank domains. Ph.D. thesis, Yale University.

Rothberg, J. M., Hartley, D., Walther, Z. & Artavanis-Tsakonas, S. (1988). *slit*: an EGF-homologous locus of *D. melanogaster* involved in the development of the embryonic central nervous system. *Cell*, **55**, 1047–1059.

Rothberg, J. M., Jacobs, J. R., Goodman, C. S. & Artavanis-Tsakonas, S. (1990). *slit*: an extracellular protein necessary for development of midline glia and commissural axon pathways contains both EGF and LRR domains. *Genes Develop.* **4**, 2169–2187.

Rupp, F., Payan, D. G., Magill-Solc, C., Cowan, D. M. & Scheller, R. H. (1991). Structure and expression of a rat agrin. *Neuron*, **6**, 811–823.

Ryseck, R. P., MacDonald, B. H., Mattei, M. G. & Bravo, R. (1991). Structure, mapping and expression of fisp-12, a growth factor inducible gene encoding a secreted cysteine-rich protein. *Cell Growth Differ.* **2**, 225–233.

Sasaki, M., Kleinman, H. K., Huber, H., Deutzmann, R. & Yamada, Y. (1988). Laminin, a multidomain protein. *J. Biol. Chem.* **263**, 16536–16544.

Simmons, D. L., Levy, D. B., Yannoni, Y. & Erikson, R. L. (1989). Identification of a phorbol ester-repressible v-src-inducible gene. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 1178–1182.

Voorberg, J., Fontijn, R., Calafat, J., Janssen, H., van Mourik, J. A. & Pannekoek, H. (1991). Assembly and routing of von Willebrand factor variants: the requirements for disulfide-linked dimerization reside within the carboxy-terminal 151 amino acids. *J. Cell. Biol.* **113**, 195–205.

Wagner, D. D., Saffaripour, S., Bonfanti, R., Sadler, J. E., Cramer, E. M., Chapman, B. & Mayadas, T. N. (1991). Induction of specific storage organelles by von Willebrand Factor propolypeptide. *Cell*, **64**, 403–413.

Edited by M. Yaniv