Protein Science (1995), *4*:1421-1425. Cambridge University Press. Printed in the USA.
Copyright © 1995 The Protein Society

FOR THE RECORD

The SEA module: A new extracellular domain associated with *O*-glycosylation

PEER BORK[1,2] AND LASZLO PATTHY[3]
[1] Max-Delbrück-Center for Molecular Medicine, 13125 Berlin-Buch, Germany
[2] European Molecular Biology Laboratory, 69117 Heidelberg, Germany
[3] Hungarian Academy of Sciences, Institute of Enzymology, Budapest, Hungary

(RECEIVED February 23, 1995; ACCEPTED April 21, 1995)

Abstract: Using a variety of homology search methods and multiple alignments, a new extracellular module was identified in (1) agrin, (2) enterokinase, (3) a 63-kDa sea urchin sperm protein, (4) perlecan, (5) the breast cancer marker MUC1 (episialin), (6) the cell surface antigen 114/A10, and (7/8) two functionally uncharacterized, probably extracellular, *Caenorhabditis elegans* proteins. Despite the functional diversity of these adhesive proteins, a common denominator seems to be their existence in heavily glycosylated environments. In addition, the better characterized proteins mentioned above contain all *O*-glycosidic-linked carbohydrates such as heparan sulfate that contribute considerably to their molecular masses. The common module might regulate or assist binding to neighboring carbohydrate moieties.

Keywords: agrin; enterokinase; homology search; molecular evolution; perlecan; sperm protein

In animals, most of the extracellular proteins are composed of multiple modules, i.e., independent building blocks that are found in functionally diverse proteins (for reviews see Doolittle, 1985; Patthy, 1985, 1991; Baron et al., 1991; Bork, 1991, 1992; Doolittle & Bork, 1993; Bork & Bairoch, 1995). The spread of these protein modules can only be the result of genetic shuffling mechanisms; the genomic organization of many of them is consistent with exon shuffling, which requires phase-compatible introns (for reviews see Patthy, 1991, 1994). In some cases, the shuffled modules retain similar functions in different proteins, but often they are only used as structural scaffolds and new functions evolve after the shuffling event. Nevertheless, deducing the modular architecture of an extracellular protein not only sheds light on its structural features, but also provides clues about its functional units and binding sites.

The recent sequencing of enterokinase, the biochemically well-characterized initiator of digestion, revealed a highly modular architecture and the enzymatic activity could be assigned to the C-terminal domain (light chain after processing), which contains a serine protease of the trypsin type (Kitamoto et al., 1994; Matsushima et al., 1994). The protease domain appears to be most similar to hepsin and blood clotting enzymes such as factor XI and prekallekrein; the modular architecture of the N-terminal heavy chain is, however, considerably different from blood clotting factors (Fig. 1).

By analogy to other modular proteins, the only two segments of enterokinase for which no homology has been found yet are also expected to form modules, i.e., they should also be present in other extracellular proteins as structurally independent building blocks. When subjecting these segments to a variety of sequence analysis methods (for details see Koonin et al., 1994), we indeed found that the N-terminus of enterokinase is similar to segments in other extracellular mosaic proteins. Thus, we report here the delineation of a new protein module and discuss structural and functional features.

Results and discussion: A combination of computer methods, including multiple Blastp database searches, as well as iterative pattern and profile approaches (for details see Materials and methods), revealed a new protein module in enterokinase, agrin, a sea urchin sperm protein, perlecan, MUC1 (episialin), the cell surface antigen 114/A10, and two functional not well-characterized *Caenorhabditis elegans* proteins (Fig. 1). We will refer to this new domain as the SEA module after the first three proteins in which it was identified (sperm protein, enterokinase and agrin).

The beginning of the module can be defined by a phase 1 intron (i.e., the intron is inserted after the first base in the codon) as determined from the genomic structures of agrin and perlecan (Rupp et al., 1992a; Cohen et al., 1993). This is consistent with the end of preceding modules (Fig. 2) in all the identified proteins (internal signal sequence in enterokinase). The sequence similarity between all proteins containing SEA modules is significant over a length of about 80 residues (Fig. 2). In all the proteins shown in Figure 1, an about 40-amino acid-long sequence segment separates the conserved 80-residue region of the SEA module from the subsequent downstream modules. This segment was included in the alignment (Fig. 2) because it proba-

Reprint requests to: Peer Bork, European Molecular Biology Laboratory, Postfach 10.2209, Meyerhofstrasse 1, 69117 Heidelberg, Germany; e-mail: bork@embl-heidelberg.de.

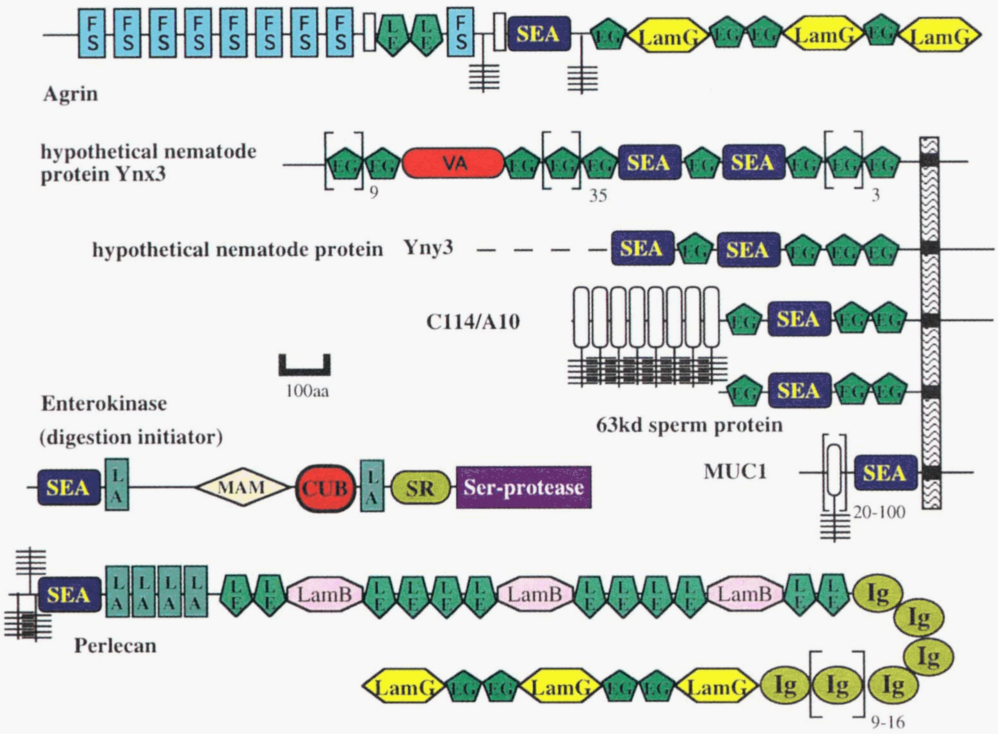

Fig. 1. Modular architecture of the proteins containing the SEA module. Names and abbreviations were chosen according a nomenclature for extracellular protein modules proposed at a recent meeting on modular proteins (for details see Bork & Bairoch, 1995). FS, follistatin-like module; LE, laminin EGF-like, found in laminin and similar matrix proteins; LamG, first identified as G domain in laminin; EG, EGF-like module; VA, Von Willebrand factor type A module; SEA, SEA module; LA, LDL receptor class A module; MAM, also found in meprin, A4 protein, and receptor protein phosphatase μ; CUB, present in a variety of developmentally regulated and C1s-like complement proteins; SR, first identified in the scavenger receptor but now known to occur in numerous other proteins. Putative *O*-glycosylation sites are denoted by antenna-like symbols. Putative transmembrane regions of the C1s and the GPI anchor of the sea urchin SP63 are indicated.

bly extends the C-terminal boundary of the SEA module (e.g., to the alternatively spliced exon in agrin and enterokinase or the phase 1 introns in perlecan; Fig. 2). This appears to be the upper boundary for the domain as other modules follow in the proteins shown in Figure 1. The distribution and phase of introns in all SEA modules (where determined) differs somewhat, although clusters of their boundaries (e.g., a common phase 0 intron within the conserved part of the alignment; Fig. 2) give additional genetic support for a homology of all SEA modules mentioned above.

Secondary structure predictions for the conserved 80 residues suggest successive β-strands interrupted by one α-helix (Fig. 2). A search with a string of secondary structure elements derived from Figure 2 in a database of selected proteins with known three-dimensional structure (L. Holm, pers. comm.) revealed the best matches with the fold of streptococcal protein G (PDB code 2GB1; Gronenborn et al., 1991), which belongs, together with ubiquitin, to family 44 of the classification of Holm and Sander (1994). In protein G, two β-hairpins are formed, respectively, by strands A and B and strands C and D (Fig. 2). The two β-hairpins are connected via hydrogen bonds between the parallel, central strands A and D. The α-helix between strands B and C crosses the central sheet stabilizing the two β-hairpins (for structural details and immunoglobulin-binding site, see Derrick & Wigley, 1994). The conserved positions in the multiple alignment of the SEA modules are consistent with the fold described above: the most hydrophobic β-strands are A and D (Fig. 2), which would be the central strands of the β-sheet. The nonconserved extension following the conšerved region is predicted to contain another helix and two β-strands (Fig. 1). This C-terminus might be an addition to the protein G like topology, but may also act as a linker region.

When speculating on the functional role of the SEA modules, it seems important to note that all the better characterized proteins shown in Figure 1 act in heavily glycosylated surroundings and are all *O*-linked proteoglycans, with the carbohydrates contributing a considerable fraction of their molecular weight. Otherwise, the overall function of the proteins containing SEA modules is diverse.

Agrin, a heparan sulfate proteoglycan of the basal lamina of the neuromuscular junction (Tsen et al., 1994) plays a key role in the formation and maintenance of the synapse as well as postsynaptic differentiation (Nastuk & Fallon, 1993; Patthy & Nikolics, 1993). It is responsible for the clustering of acetylcholine receptors (AChRs) and other proteins at the neuromuscular junction. For these functions, it interacts with several proteins including AChR, nidogen, and α-dystroglycan; binding to the latter triggers the aggregation of AChRs and heparan sulfate proteoglucan AChE in the postsynaptic membrane (Fallon & Hall, 1994, and references therein). The C-terminal third of agrin seems to bind dystroglycan (Fallon & Hall, 1994, and references therein), but several other interactions are needed

```
intron phase                  1 1     A              B                  0 0                   C   2 2            D     1
secondary struct.             LL.eEEEEEEEEE.LL..eEee.L.LLLL.HHHHHHHHHHHHHHHHHH..LLLLL.eEEEEEEEeLLLLL.....eeeeEEEee..

Ynx3_Caeel-2  2333   VVESWNVPLWVVRDKEKPIVFSESFDNPQTPVYKDYSKRLEKGIEGCYPHTELKN-AFVTAEVNDIVNPVLMNASYDTGLLFNTTV
Yny3/Caeel-2     ?   AVESWNLPLYVIRDGHEKITYSPSLSNPLNDDHKDLVSRFESGVAQSYDKTPLKG-AFVTAEVNEIENPESRKKSWDTGILYNFTS
Ynx3_Caeel-1  2156   EVQETPFELRVVTRDQRPLMYSTEFGSQKSPSYVEIVELFEkNMARTFGGTSLAP-RYVNTKVDYITHPKTKNSSWDQGLLFKYEV
Yny3/Caeel-1     ?   PTTSIPLVVRVMEYDGEPIQYRTDYSKPDTQAHIEIVDAVKkSVGKIIGKTDVAP-RFVTTDVNYITNPKVQNSEWDKGLLGNVSV
Entp/Pig        51   LGKSHEARGTMKITSG--VTYNPNLQDKLSVDFKVLAFDIQQMIGEIFQSSNLKN-EYKNSRVLQFENGSVI---VIFDLLFAQWV
Entp/Bovin      51   FGKSHEARGTLKIISG--ATYNPHLQDKLSVDFKVLAFDIQQMIDDIFQSSNLKN-EYKNSRVLQFENGSII---VIFDLLFDQWV
Agri_Rat      1023   ATKAFQGVLELEGVEGQELFYTPEMADPKSELFGETARSIESTLDDLFRNSDVKK-DFWSVRLRELGPGKLVR--AIVDVHFDPTT
Agri/Mouse    1023   aTKAFQGALELEGVEGQELFYTPELADPKSDLFRETARSIESTvDDLFRNSDVKK-DFWTIRLRELGPGKLVR--AIVDVHFDPAt
Agri_Chick    1026   ATKVFQGVLILEEVEGQELFYTPEMADPKSELFGETARSIESALDELFRNSDVKN-DFKSIRVRDLGQSSAVR--VIVESHFDPAT
Agri/Disom     411   PTKLFQGVLIVEEVEGQELFYTPEMDDPKSELFGETARSIENALNELFGNSNVKK-DFKSVRVHGLGPSDPVR--IIVEVHFDPRT
Sp63/Strpu      81   VAQQFAGSFSVTQVGGSNVLYSADLADTDSAAFASLAADVEDALDTVYQASTMAD-IYLGSEVWGV-PEWLYR--GRLHVLFATED
Muc1_Mouse     411   PQLSVGVSFFFLFFYIQNHPFNSSLEDPSSNYYQELKRNISGLfLQIFNG------DFLGISSIKFrSGSVV---VESTVVFREGT
Muc1_Human    1034   PQLSTGVSFFFLSFHISNLQFNSSLEDPSTDYYQELQRDISEMfLQIYKQG-----GFLGLSNIKFrPGSVV---VQLTLAFREGT
Perl/Human      80   QMvYFRALVNFTR----SIEYSPQLEDAGSREFREVSEAVVDTlESEYLKIPGD--QVVSVVFIkELDGWVF---VELDVGSEGNA
Perl/Mouse      80   QMVYFRALVNFTR----SIEYSPQLEDASAKEFREVSEAVVEKLEPEYRKIPGD--QIVSVVFIKELDGWVF---VELDVGSEGNA
C114_Mouse     272   CVKGTTFPGDISM----SVSETANLEDENSVGYQELYNSVTDFFETTFNKTDYGQTVIIKVSTAPSRSARSAMRDATKDVSVSVVN

consensus                hth h h        h attthttt o   a   h tth t h t att        hh         tt   h    h hthh t

intron phase                                          0       01 1
secondary struct.    ..LL.hhhHHHHHHHhh.L..eeeeee.L.1..eeeee..L

Ynx3_Caee2           HFRKGMVHVPSDAYYQLIK--YVTKENNNEVGDSELYLNPT   2458   Z30423
Yny3/Caee2           HFVKGSVAEPASVFTDLID--YIQKRNDFEvGKSKLFISPE      ?   L16679
Ynx3_Caeel           QTTKQSQPIDECELWKQMQ--ASLDRTNGAIGGGSLRVASD   2283   Z30423
Yny3/Caeel           H-LAGKEEVDKCRFYEQFA--EIVREMGGRVDRIKLSDDAD      ?   L16679
Entp/Pig             --SDENIKEELIQGIEANKS-SQLVAFHIDVNSIDITESLE    170   D30799
Entp/Bovin           --SDKNVKEELIQGIEANKS-SQLVTFHIDLNSIDITaSLE    170   U09859
Agri_Rat             AFQASDVGQALLRQIQVSRP-WALAVRRPLQEHVRFLD-FD   1149   P25304
Agri/Mouse           AFQAPDVGQALLQQIQVSRP-WALAVRRPLPEHVRFLD-Fd   1149   M92657
Agri_Chick           SYTAADVQAASLKQIRASKK-RTILVKKPQQEHVKFMD-FD   1149   P31696
Agri/Disom           SYNSHDVQRALLQQVKQSRR-KSIVVKKPEQDNVKIVD-FD    533   L01423
Sp63/Strpu           --AGQPVLVNSTDATE-----AFTTALAAEAANLGITI-DD    196   M99584
Muc1_Mouse           -FSASDVKSQLIQHKKEADS-YNLTISEVKvNEMQFPPSAQ    527   Q02496
Muc1_Human           -INVHDVETQFNQYKTEAASRYNLTISDVSvSDVPFPFSAQ   1152   P15941
Perl/Human           --DGAQIQEMLLRVISSGSV-ASYVTSPQGFQFRRLGTvPQ    195   L22078
Perl/Mouse           --DGSQIQEVLHTVVSSGSI-GPYVTSPWGFKFRRLGTVPQ    195   M77174
C114_Mouse           -IFGADTKETEKSVSSAIE--TAIKTSGNVKDYVSINL---    390   P19467

consensus                tttht    t ht        h    t    t   th      t
```

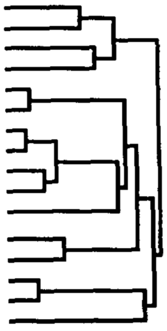

Fig. 2. Alignment and dendrogram of the SEA module. First column, protein name/species abbreviations (SWISS-PROT database codes with an underscore in the name are used if available: Ynx3 and Yny3, *C. elegans* proteins; Entp, enterokinase; Agri, agrin; Sp63, 63-kDa sperm protein; Muc1, MUC1; Perl, perlecan; C114, 114/A10); second column, position of the module in the respective proteins; last two columns, position of the end of the alignment and database accession numbers. Top line shows intron phase as deduced from the corresponding genomic sequences (e.g., 0 corresponds to intron insertions before the affected amino acid; 1 indicates intron between the first and the second base of the codon; the respective amino acids are shown as lowercase letters and italics in the alignment), A–D correspond to the β-strands mentioned in the text; second line, predicted secondary structure by using the PHD program (Rost & Sander, 1994): E(e), β-strand; H(h), helix; L (not assigned), not helix or β-strand. Capitals denote positions with an expected accuracy larger than 82% and lowercase letters or unassigned positions have less predictive power; average accuracy of the method is greater than 72% (Rost & Sander, 1994). Because the lower part of the alignment (putative linker region) is not significant, secondary structure predictions are omitted for these segments. Bottom consensus line: capital letters, amino acids conserved in more than 50% of the sequences; t, turn-like or polar; h, hydrophobic; a, aromatic; o, S or T. Using ClustalW (Thompson et al., 1994), dendrograms for both the conserved segment (upper part) and the whole region as displayed were carried out and did not differ significantly. Shown is the result for the complete alignment.

for proper functioning. The N-terminal part is essential for association with the extracellular matrix and/or cell surface molecules (Ferns et al., 1993); the two cysteine-rich so-called EGF-like laminin modules (Fig. 1) are assumed to bind nidogen as they do in laminin (Nastuk & Fallon, 1993). Thus, the SEA module might also interact with components of the basal lamina or cell surface molecules. A specific binding function is supported by the high conservation of the SEA module in agrins from different species (Tsim et al., 1992).

The digestion initiator enterokinase is physiologically the only enzyme that converts trypsinogen into trypsin in pancreatic fluid (Light & Janska, 1989). It is thought to interact with the intestinal brush border membrane through its modular noncatalytic heavy chain. By analogy with the complex regulatory proteins in the coagulation, fibrinolytic, and complement activation cascades (Patthy, 1993), it seems likely that the various modules of enterokinase (Fig. 1) also mediate interactions with cell surface-associated macromolecules and thus regulate enterokinase activiy. Enterokinase is heavily glycosylated and probably contains 30–50% carbohydrate (Kitamoto et al., 1994).

The 63-kDa sea urchin sperm protein (SP63) has been suggested to be a receptor for egg jelly ligands triggering the sperm acrosome reaction, but other experiments have shown that it is not a speract receptor (Mendoza et al., 1993). Because sperm exocytosis in animals involves interactions with heavily glycosylated proteins of the zona pellucida, an extracellular matrix surrounding oocytes (for review see Wassarman, 1988), carbohydrate binding may also be needed for the function of the 63-kDa sperm protein. Although its overall function remains unknown, the structural similarity to developmentally regulated transmembrane proteins suggests that it might mediate sperm–egg or sperm–matrix interactions (Mendoza et al., 1993).
]

The cell surface antigen 114/A10 has a similar overall architecture to the sea urchin protein SP63 (Fig. 1) but is an integral transmembrane protein (compared to the glycosylphosphatidyl inositol [GPI]-anchored SP63) and has an N-terminal extension with eight tandem repeats containing *O*-linked carbohydrates that have a molecular weight three to five times higher that the amino acid composition alone (Dougherty et al., 1989). The heavily glycosylated 114/A10 is highly expressed in hemopoietic progenitor cells and IL-3-dependent cell lines, and thus it has been proposed to play a regulatory role in cellular responses to IL-3 (Dougherty et al., 1989).

Perlecan is a component of all basement membranes and pericellular matrices. It contains three *O*-linked heparan sulfate chains, their attachment sites in the core protein just precede the SEA module (Fig. 2). Human and mouse genomes contain a single perlecan gene that is subjected to alternative splicing. Perlecan has been proposed to bind to various extracellular matrix and cell surface proteins (for review see Iozzo et al., 1994). Its very early expression in development and the ability to promote the binding of basic fibroblast growth factor to its receptor suggest a role in cell and tissue growth (Aviezer et al., 1994).

MUC1 stands for a group of alternatively spliced, polymorphic, mucin-like glycoproteins that have a high molecular mass due to their *O*-linked carbohydrates (Spicer et al., 1991; Zrihan-Licht et al., 1994). MUC1 proteins are expressed at basal levels by most secretory epithelial cells, but their expression is dramatically increased in malignant breast epithelia and they are thus important breast cancer markers. The expression of MUC1 may reduce cellular adhesion (Zrihan-Licht et al., 1994, and references therein).

Little is known about the function of the two putative nematode proteins that were identified within the *C. elegans* genome sequencing project (Wilson et al., 1994). Like other multiple EGF domains containing integral membrane proteins, they may also participate in cell–cell or cell–matrix interactions.

When comparing orthologues from different species, the SEA modules always belong to the most conserved regions in the respective proteins, which is suggestive of a functional domain. The most striking common feature of all the proteins mentioned above seems to be, however, the coexistence of *O*-glycosidic-linked carbohydrates and SEA modules in four out of the eight distinct proteins (Fig. 1). Two others (SP63 and enterokinase) are known to be heavily glycosylated; the remaining two *C. elegans* proteins have not yet been studied. All these proteins contain several consensus motifs that are apparently required for *O*-glycosylation and could therefore also be *O*-linked proteoglycans. Furthermore, Unc52 from *C. elegans* (Rogalski et al., 1993) with a modular architecture similar to perlecan and grouped into the perlecan family (Iozzo et al., 1994) lacks both the N-terminal *O*-glycosylation sites and the succeeding SEA module. Splice variants of MUC1 devoid of the tandem repeats that contain the carbohydrate attachment sites contain only a short probably nonfunctional segment of the SEA module (Zrihan-Licht et al., 1994).

It needs to be experimentally verified whether all of the SEA module-containing proteins are indeed *O*-glycosidic-linked proteoglycans as is suggested by our comparisons; at least they seem to function in carbohydrate-rich environments. Furthermore, all of the characterized proteins containing SEA modules interact with constituents of the extracellular matrix. The functional role of the SEA module remains unclear, it might even assist in recognizing the attachment sites to be glycosylated during the posttranslational modification process. A precise description of the binding function of SEA modules is also impossible at the moment; only structurally important residues seem to be conserved (Fig. 2), and individual SEA modules may well have distinct binding targets in different proteins.

Materials and methods: The initial database searches were carried out with the N-terminal 118-residue-long segment of the mature peptide preceding an alternatively spliced exon and the first identified module (Fig. 1; Kitamoto et al., 1994). Using the programs of the Blast series and applying several amino acid substitution matrices (Altschul et al., 1990), the best scoring database proteins identified were agrins from different species (Ruegg et al., 1992; Rupp et al., 1992b; Smith et al., 1992; Tsim et al., 1992) and a 63-kDa sea urchin sperm protein (Mendoza et al., 1993). They had probability (*P*) values of matching by chance (Altschul et al., 1990) between 0.004 and 0.14. Blastp *P*-values in this range are not low enough to infer common ancestry immediately, but are often indicative of a distant homology (that has to be verified by other methods). Curiously, the Blastp alignments of the detected segments were consistent, i.e., the same region of the enterokinase segment was matched and similar amino acids were conserved. This prompted further studies, and a multiple sequence alignment using the program ClustalW (Thompson et al., 1994) was carried out. Both enterokinases from cow (Kitamoto et al., 1994) and pig (Matsushima et al., 1994), as well as agrins from rat, mouse, chicken, and marine ray were included. The alignment (training set) was used for several pattern and profile searches (Gribskov et al., 1987; Patthy, 1987; Rohde & Bork, 1993; Tatusov et al., 1994) that worked complementarily for this protein family. Each method was trained with both the whole alignment and with conserved regions. If any of the methods mentioned above significantly identified a putative new member of the family, blast searches with the new member were carried out, the multiple alignment was reconstructed, and the new member was thus included in the training set. The closest similarity had the two repeats in the *C. elegans* proteins as all of the profile and pattern methods which are able to identify internal repeats recognized them first. MUC1, perlecan, and 114/A10 were added in additional iterations of the procedure.

Finally, the MoST program (Tatusov et al., 1994) was used to check the significance of detected similarities. When subjecting the conserved regions of the final alignment (Fig. 2) to database searches, all proteins of the family had *P*-values of matching by chance below 10^{-5}, whereas other proteins of the database scored above 0.01, i.e., a clear discrimination was achieved.

For secondary structure prediction, the neural network method PHD (Rost & Sander, 1994) was used with the ClustalW alignment (Thompson et al., 1994) as input. The topology prediction was based on the result of a filtering procedure that extracted three-dimensional structures deposited in PDB (Bernstein et al., 1977) with a similar order of secondary structure elements (L. Holm, unpubl.).

Acknowledgments: We thank Liisa Holm for help with the topology prediction and Rebecca Wade for critical reading of the manuscript.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman D. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

Aviezer D, Hecht D, Safran M, Eisinger M, David G, Yayon A. 1994. Perlecan, basal lamina proteoglycan, promotes basic fibroblast growth factor-receptor binding, mitogenesis, and angiogenesis. *Cell 79*:1005–1013.

Baron M, Norman DG, Campbell ID. 1991. Protein modules. *Trends Biochem Sci 16*:13–17.

Bernstein FC, Koetzle TF, Williams GJB, Meyers EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

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, Bairoch A. 1995. A proposed nomenclature for the extracellular protein modules of animals. *Trends Biochem Sci 20*(3):poster C02.

Cohen IR, Grässel S, Murdoch AD, Iozzo RV. 1993. Structural characterization of the complete human perlecan gene and its promotor. *Proc Natl Acad Sci USA 90*:10404–10408.

Derrick JP, Wigley DB. 1994. The third IgG-binding domain from streptococcal protein G. *J Mol Biol 243*:906–918.

Doolittle RF. 1985. The genealogy of some recent evolved vertebrate proteins. *Trends Biochem Sci 10*:233–237.

Doolittle RF, Bork P. 1993. Evolutionarily mobile modules. *Sci Am 269*(4):50–56.

Dougherty GJ, Kay RJ, Humphries RK. 1989. Molecular cloning of 114/A10, a cell surface antigen containing highly conserved repeated elements, which is expressed by murine hemopoietic progenitor cells and interleukin-3 dependent cell-lines. *J Biol Chem 264*:6509–6514.

Fallon RR, Hall ZW. 1994. Building synapses: Agrin and dystroglycan stick together. *Trends Neurosci 17*:469–473.

Ferns M, Campanelli JT, Hoch W, Scheller RH, Hall ZW. 1993. The ability of agrin to cluster AChRs depends on alternative splicing and on cell surface proteoglucans. *Neuron 11*:491–502.

Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: Detection of distant similarities. *Proc Natl Acad Sci USA 84*:4355–4358.

Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT, Clore GM. 1991. A novel, highly stable fold of the immunoglobulin-binding domain of streptococcal protein G. *Science 253*:657–661.

Holm L, Sander C. 1994. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res 22*:3600–3609.

Iozzo RV, Cohen IR, Grässel S, Murdoch AD. 1994. The biology of perlecan: The multifaceted heparan sulphate proteoglycan of basement membranes and pericellular matrices. *Biochem J 302*:625–639.

Kitamoto Y, Yuan X, Wu Q, McCourt DW, Sadler JE. 1994. Enterokinase, the inhibitor of intestinal digestion, is a mosaic protease composed of a distinctive assortment of domains. *Proc Natl Acad Sci USA 91*:7588–7592.

Koonin EV, Bork P, Sander C. 1994. Yeast chromosome III: New gene functions. *EMBO J 13*:493–503.

Light A, Janska H. 1989. Enterokinase (enteropeptidase): Comparative aspects. *Trends Biochem Sci 14*:110–112.

Matsushima M, Ichinose M, Yahagi N, Kakei N, Tsukada S, Miki K, Kurokawa K, Tashiro K, Shiokawa K, Shinomiya K, Umeyama H, Inoue H, Takahashi T, Takahashi K. 1994. Structural characterization of porcine enterokinase. *J Biol Chem 269*:19976–19982.

Mendoza LM, Nishioka D, Vacquier VD. 1993. A GPI-anchored sea urchin membrane protein containing EGF domains is related to uromodulin. *J Cell Biol 121*:1291–1297.

Nastuk MA, Fallon JR. 1993. Agrin and the molecular choreography of the synapse formation. *Trends Neurosci 16*:72–76.

Patthy L. 1985. Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. *Cell 41*:657–663.

Patthy L. 1987. Detecting homology of distantly related proteins with consensus sequences. *J Mol Biol 198*:567–577.

Patthy L. 1991. Modular exchange principles in proteins. *Curr Opin Struct Biol 1*:351–361.

Patthy L. 1993. Modular design of proteases of coagulation, fibrinolysis and complement activation: Implications for protein engineering and structure-function studies. *Methods Enzymol 222*:10–21.

Patthy L. 1994. Introns and exons. *Curr Opin Struct Biol 4*:404–412.

Patthy L, Nikolics K. 1993. Function of agrins and agrin-related proteins. *Trends Neurosci 16*:76–81.

Rogalski TM, Williams BD, Mullen GP, Moerman DG. 1993. Products of the unc-52 gene in *C. elegans* are homologous to the core protein of the mammalian basement membrane heparan sulfate proteoglycan. *Genes & Dev 7*:1471–1484.

Rohde K, Bork P. 1993. A fast, sensitive pattern-matching approach for protein sequences. *CABIOS 9*:183–189.

Rost B, Sander C. 1994. Combining evolutionary information and neural networks to predict secondary structure. *Proteins Struct Funct Genet 19*:55–72.

Ruegg MA, Tsim KWK, Horton SE, Kroger S, Escher G, Gensch EM, McMahan UJ. 1992. The agrin gene codes for a family of basal lamina proteins that differ in function and distribution. *Neuron 8*:691–699.

Rupp F, Ozcelik T, Linial M, Peterson K, Francke U, Scheller R. 1992a. Structure and chromosomal localization of the mammalian agrin gene. *J Neurosci 12*:3535–3544.

Rupp F, Payan DG, Magill-Solc C, Cowan DM, Scheller RH. 1992b. Structure and expression of rat agrin. *Neuron 6*:811–823.

Smith MA, Magill-Solc C, Rupp F, Yao YMM, Schilling JW, Snow P, McMahan UJ. 1992. Isolation and characterisation of a cDNA that encodes an agrin homolog in the marine ray. *Mol Cell Neurosci 3*:406–417.

Spicer AP, Parry G, Patton S, Gendler SJ. 1991. Molecular cloning and analysis of the mouse homologue of the tumor-associated mucin, MUC1, reveals conservation of potential *O*-glycosylation sites, transmembrane and cytoplasmic domains and loss of minisatellite-like polymorphism. *J Biol Chem 266*:15099–15109.

Tatusov R, Altschul SF, Koonin EV. 1994. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci USA 91*:12091–12095.

Thompson JD, Higgins DG, Gibson T. 1994. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res 22*:4673–4680.

Tsen G, Halfter W, Kroger S, Cole G. 1994. Agrin is a heparan sulfate proteoglucan. *J Biol Chem 270*:3392–3399.

Tsim KWK, Ruegg MA, Escher G, Kroger S, McMahan UJ. 1992. cDNA that encodes active agrin. *Neuron 8*:677–689.

Wassarman PM. 1988. Zona pellucida glycoproteins. *Annu Rev Biochem 57*:415–442.

Wilson R, Ainscough R, Anderson K, Baynes C, Berks M, Bonfield J, et al. 1994. 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature 368*:32–38.

Zrihan-Licht S, Vos HL, Baruch A, Elroy-Stein O, Sagiv D, Keydar I, Hilkins J, Wreschner DH. 1994. Characterisation and molecular cloning of a novel MUC1 protein, devoid of tandem repeats, expressed in human breast cancer tissues. *Eur J Biochem 224*:787–795.