J. Mol. Biol. (1998) 275, 725–730

JMB

COMMUNICATION

Merging Extracellular Domains: Fold Prediction for Laminin G-like and Amino-terminal Thrombospondin-like Modules Based on Homology to Pentraxins

Georg Beckmann[1]*, Jens Hanke[1], Peer Bork[1,2] and Jens G. Reich[1]

[1]*Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Str.10, 13122 Berlin-Buch, Germany*

[2]*European Molecular Biology Laboratory, Meyerhofstr. 1 69012 Heidelberg, Germany*

**Corresponding author*

Using a new method for construction and database searches of sequence consensus strings, we have identified a new superfamily of protein modules comprising laminin G, thrombospondin N and the pentraxin families. The conserved patterns correspond mainly to hydrophobic core residues located in central beta strands of the known three-dimensional structures of two pentraxins, the human C-reactive protein and the serum amyloid P-component. Thus, we predict a similar jellyroll fold for all members of this superfamily. In addition, the conservation of two exposed aspartate residues in the majority of superfamily members suggests hitherto unrecognised functional sites.

Keywords: sequence analysis; protein modules; laminins; thrombospondins; pentraxins

Shuffled protein modules appear to be the building stones of the majority of extracellular proteins in animals (Doolittle, 1995; Bork & Bairoch, 1995; Patthy, 1996). For more than half of the about 70 described extracellular protein modules, the three-dimensional (3D) structure has been determined (Bork *et al.*, 1996). Sometimes the 3D structures revealed surprising relations between module families (e.g. the link and the C-type lectin modules; Kohda *et al.*, 1996) that allow the evolution of those modules to be traced and an understanding of some functional properties.

Here, we merge three module families, namely laminin G (LamG), amino-terminal thrombospondin (TspN) and the pentraxins, into a novel superfamily. Their relationship extends to about 140 residues containing patterns of alternating hydrophobicity characteristic of antiparallel beta-strands. The module has obviously diverged during evolution, so that the sequence similarity has receded into the weak ''twilight zone''. Our evidence is based on very sensitive fine-tuned sequence analysis. As 3D-structures and a theoretical model are available for pentraxins (Emsley *et al.*, 1994; Shrive *et al.*, 1996; Srinivasan *et al.*, 1994), we are able to infer a tentative structure prediction for the common domain.

LamG and TspN-containing proteins belong to the extensive class of multi-domain adhesive proteins in the matrix that surrounds animal cells. They act as molecular bridges between cells and matrix, and participate, due to their numerous binding sites, in cell-cell communication.

The LamG domain was originally identified as fivefold repetition of about 158 to 180 residues in the C-terminal globular (hence G) domain of the laminin alpha-1 chain (Deutzmann *et al.*, 1988). The LamG family comprises a multitude of diverse proteins (reviewed by Patthy, 1991, 1992; Joseph & Baker, 1992; Rothberg & Artavanis-Tsakonas, 1992; Ushkaryov *et al.*, 1992; Manfioletti *et al.*, 1993). A host of binding functions has been ascribed to LamG modules (e.g. see Yurchenko *et al.*, 1993; Bocchinfuso & Hammond, 1994; Brancaccio *et al.*, 1995; Mercurio, 1995; Mark *et al.*, 1996; Delwel & Sonnenberg, 1996), which points to a multifarious role in cell adhesion, signalling, migration, assembly and differentiation.

Thrombospondins are likewise secreted multimodular glycoproteins to which a bewildering diversity of functions has been assigned (for reviews, see Adams & Lawler, 1993; Lawler *et al.*, 1993; Bornstein & Sage, 1994; Bornstein, 1995). These molecules start at their amino terminus with a short signal sequence, followed by TspN, which

Abbreviations used: LamG, laminin G; TspN, amino-terminal thrombospondin; CRP, C-reactive protein; SAP, serum amyloid protein.

0022–2836/98/050725–06 $25.00/0/mb971510

will be considered here; a globular module of about 210 residues that has also been found in the non-collagenous region of a variety of collagens (Bork, 1992; Mayne & Brewton, 1993).

Pentraxins (Osmand *et al.*, 1977) are a family of pentamers in cyclic symmetry consisting of subunits of about 200 amino acid residues. The most prominent members are C-reactive proteins (CRP) and serum amyloid protein (SAP). The modular nature of pentraxins is suggested by the identification of larger proteins (for a review, see Goodman *et al.*, 1996), whose C-terminal halves display clear similarity to the pentraxin domain. The ''stand-alone'' version of the pentraxins appears to have a role mainly in natural defense (complement system, phagocytosis, acute phase response to tissue infection or injury); the involvement of SAP with pathological conditions as e.g. Alzheimer's disease and diabetes mellitus and the use of CRP as an objective marker of disease activity has attracted much interest in these proteins (for a review, see Gewurz *et al.*, 1995).

The scheme in Figure 1 summarizes the relationship between the three families: (i) TspN and LamG modules form two compact clusters (ellipsoids) in sequence space and (ii) both clusters are neighboring so that a common super cluster (rectangular box) comprises both but not alien sequences (except some pentraxins, which we assert to pertain to this superfamily). This topology is proved by demonstrating two ''spliced'' consensus strings (big T and big L in Figure 1), each of which has the following property. When it serves as query in a search, the BLASTP method (Altschul *et al.*, 1990, 1997) produces a list whose highest-scoring hits consist exclusively of pertinent family members. Only at considerably lower scores (symbolized by being beyond the contour of the ellipsoid) do other sequences in the BLASTP list appear. The fact that the low-scoring matches are symmetrical for the two families in question is a strong argument for the existence of a protein superfamily with sequence similarity in the otherwise dubious twilight zone.

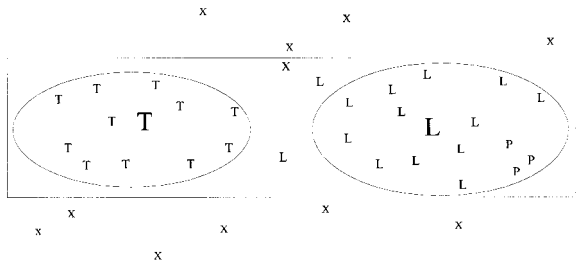

Figure 1. A drawing of the TspN/LamG/pentraxin superfamily in the sequence space. Big letters (T for TspN, L for LamG, P for pentraxin) stand for a ''consensus string'' capable of extracting by a BLASTP search (Altschul *et al.*, 1990, 1997) with cutoff value (ellipse) the pertinent family members (small letters). Left ellipse, TspNs (see Figure 2); right ellipse, LamG family with *L.polyphemus* C-reactive proteins (P for pentraxin). Some L are outside the ellipse (not within the BLASTP cutoff, see Figure 2). Box (comprising all T and L, and *Limulus polyphemus* C-reactive proteins, the TspN/LamG/pentraxin superfamily, while false hits (X) are outside the box.

We have developed a software tool called IDSECS (for **I**terative **D**atabase **S**earch with **E**volving **C**onsensus **S**trings) that allows a string that fulfils the above-mentioned criterion to be found automatically. This has been tested for more than 500 protein families, and the details of the algorithm will be published elsewhere (IDSECS is currently available on request to the authors, the associated Table of Consensus Strings, TACOS, can be queried at http://www.bioinf.mdc-berlin.de/idsecs.html). The procedure consists of an iterative improvement of an approximation to the desired consensus. Any member of the family serves as starting consensus. We use BLASTP (Altschul *et al.*, 1990) exclusively as a prefilter to reduce the number of potential candidate members. All sequences that are reported by BLASTP, regardless of their chance probabilities or scores, are taken as a new set of potential relatives. These, because they are reduced in number, are accessible to a more rigorous procedure (Smith & Waterman, 1981) that serves to select sequences according to a predefined threshold function. Given a set of selected sequences and a reference sequence, we model the consensus string on the sequences of the set. Here, pairwise alignments to the reference are used to derive a preliminary consensus that, in turn, is realigned to the sequences in question, with the new alignments yielding a new consensus, and so forth until convergence of the consensus is obtained. Hence, we coined it evolving consensus strings. It is here that IDSECS gains most of its performance. The converged consensus now enters the next iteration, which restarts by prefiltering the master database. The final consensus is obtained when in two subsequent iterations no additional sequence is selected (by the Smith-Waterman algorithm).

The final result is in most cases an oscillation around a few very similar consensus strings, which in BLASTP runs (Altschul *et al.*, 1990, 1997) ''quote'' the same set of database entries and are in turn confirmed by them as their ''representative''.

Figure 2 shows a histogram of the results of the new Gapped BLASTP search (Altschul *et al.*, 1997) with the TspN consensus string as query. The 78 top-ranking sequences are exclusively TspN-containing sequences with significant *E*-values (less than 0.001). At considerably lower scores follow more than 60 members of the LamG-family, i.e. nearly half of its members. Only ten non-members were reported intermixed at the end of the list (although the Expect parameter used, $E = 100$, awaits on average 100 false negatives). Note the logical conformity of this finding with the sketch in Figure 1.

The histogram of the BLASTP result in Figure 3 corresponds to the LamG consensus (displayed in Figures 4 and 5) as query. In addition to

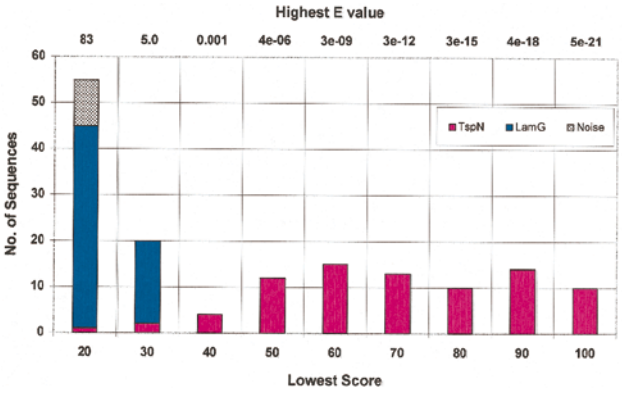

Figure 2. Retrieval of TspN-family members by a consensus string. The thrombospondin ''spliced consensus'' module (TspN) of 140 letters (displayed in Figures 4 and 5) was subjected to the new Gapped BLASTP search (Altschul *et al.*, 1997) through the NCBI non-redundant protein data base. Shown is a histogram of the BLASTP results. The *X*-axes (lowest score, highest *E*-values) denote lower boundaries of the histogram intervals. The top-ranking 78 items were exclusively members of the thrombospondin family. Sequences following at insignificant *E*-values (>0.001) comprise almost half the members of the LamG family. The original BLASTP output may be inspected under http://www.bioinf.mdc-berlin.de/pub/TspN_BLAST.html.

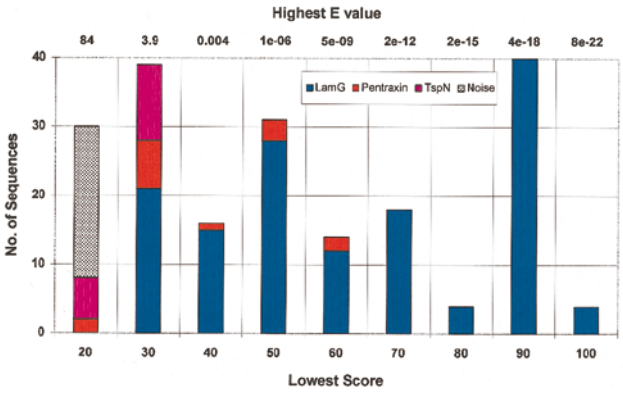

Figure 3. Retrieval of the LamG-family members by a consensus string. The laminin G consensus module (LamG) of 142 letters (displayed in Figures 4 and 5) was subjected to the new Gapped BLASTP (Altschul *et al.*, 1997) shown is a histogram of the BLASTP result. The *X*-axes (lowest score, highest *E*-values) denote lower boundaries of the histogram intervals. The top-ranking, significant (*E* value < 0.004) matches comprise members of the LamG family and six members of the pentraxin family. Amongst the latter was the C-reactive protein 1.1 from *L. polyphemus* for which a theoretical model of molecular structure exists in the PDB (accession code 1LIM; Bernstein *et al.*, 1977). The original BLASTP output is available under http://www.bioinf.mdc-berlin.de/pub/LamG_BLAST.html.

111 LamG-family members, the top-ranking sequences with significant *E*-values (less than 0.004) encompass also six members of the pentraxin family, of which the *Limulus polyphemus* C-reactive protein 1.1 was reported with an *E*-value of 3e-09, which is undoubtedly significant. Following this is a mixture of members of all three families, again accumulating several non-members at the end.

Both tables together suggest, with their reciprocal quotation, a neighborhood of TspN, LamG and some pentraxins in the sequence space and thus their affiliation to a common superfamily.

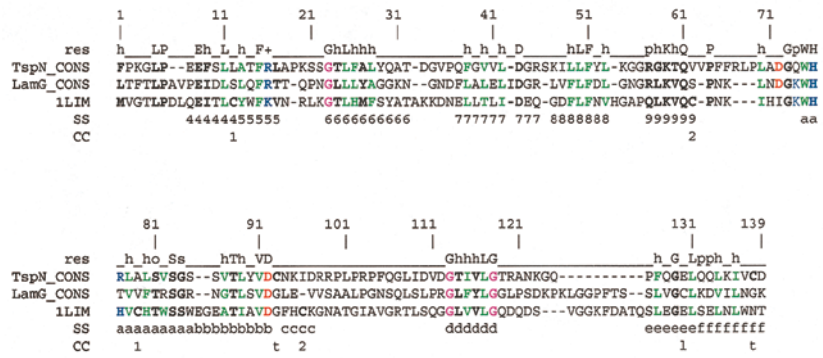

Figure 4. Alignment of consensus strings for amino-terminal thrombospondin (TspN) and laminin G-like (LamG) modules against the sequence of C-reactive protein chain 1.1 of *L. polyphemus* showing conserved region of the superfamily. TspN_CONS, LamG_CONS, amino-terminal thrombospondin (collagen) module and laminin G-like module contained as a repeated segment in the globular part of the C terminus of laminin A chain, represented here by a spliced consensus string of the most conserved regions. The representative character of the two spliced consensus strings was demonstrated in Figures 2 and 3 by their effect to extract only pertinent family members. Explanation of additional lines: 1LIM, C-reactive protein 1.1 from *L. polyphemus*, for which a structural model is contained in PDB (Bernstein *et al.*, 1977; Srinivasan *et al.*, 1994). Here, the aligned segment is shown together with salient features of its molecular structure. SS, indicates β-structure assignments in 1LIM (strands 4 through 15; two-digit numbers 10 trough 14 being replaced for convenience by a through f). Note that these features largely coincide with the most conserved parts of the sequence alignment (bold face and colored letters). CC, indicates the known or proposed cysteine bridges of 1LIM (1 and 2), of thrombospondin (t) and laminin (l, only left-hand C of this bridge situated within the alignment). Res, indicates residue conservation between the three sequences (amino acid capital letters and: h, hydrophobic; p, polar; o, serine or threonine; s, small residue; +, positively charged residue). A dash indicates a gap in the alignment. Color emphasis of residues: green, hydrophobic; magenta, conserved glycine; blue, conserved H, R and K; red, conserved D and E.

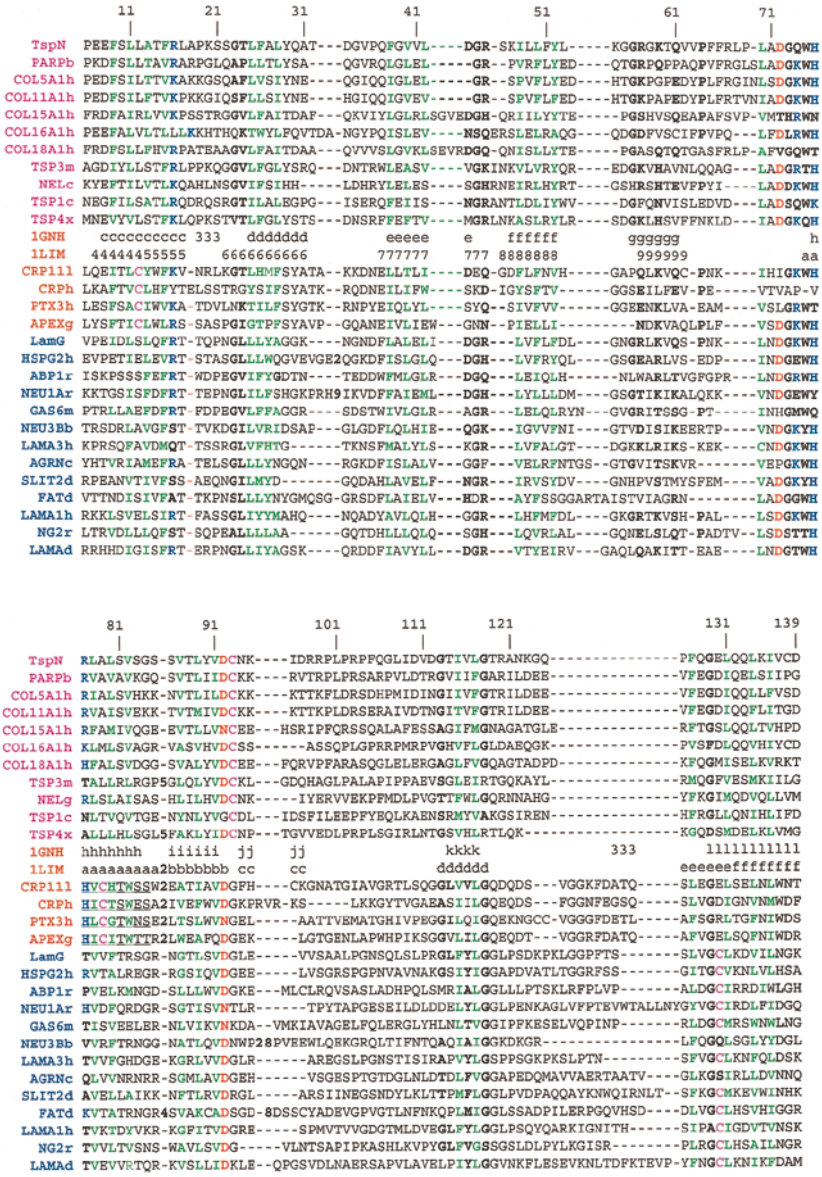

Figure 5. Multiple alignment of the common conserved region of the proposed TspN/LamG/pentraxin superfamily. Membership of sequences can be perceived through colored first column (magenta for TspN consensus (TspN) and representatives, red for pentraxins and blue for LamG consensus and samples) and by other characteristics, e.g. Cys in positions (numbering according to TspN) 12 and 79 of pentraxins, in 131 of LamGs and in 93 of TspNs. The pentraxin signature as appearing in the PROSITE database (Bairoch *et al.*, 1997) is underlined. Residue conservation is also emphasized by colored letters in the alignment: green, hydrophobic; blue, positively charged (H, K and R); red, aspartate or asparagine (D, N); purple, cysteine residues; bold-face black, largely conserved other residues. Note that residue colors reside mainly in conserved β-strands, which are shown in the 1GNH (3D-structure, Shrive *et al.*, 1996) and 1LIM (theoretical model, Srinivasan *et al.*, 1994) lines (codes from the Protein Data Bank, Bernstein *et al.*, 1977). Other acronyms in the first column are (last letter denotes organism: b, bovine; c, chicken; d, *Drosophila*; g, guinea pig; h, human; l, *Limulus polyphemus*; m, mouse; r, rat; x, *Xenopus laevis*): PARP, proline and arginine-rich protein; COLxAy, alpha y chain of collagen x; TSP, thrombospondins; CRP11, C-reactive protein 1.1; PTX3, pentraxin-like protein PTX3; APEX, acrosomal pentraxin-like protein apexin; HSPG2, heparan sulphate proteoglycan HSPG2; ABP1, androgen-binding protein 1; NEU1A, 3B, neurexin 1 alpha, 3 beta; LAMA(1,3), laminin alpha (1,3) chains; AGRN, agrin; FAT, cadherin-related tumor suppressor protein FAT; SLIT2, neurogenic locus protein SLIT-2; NG2, chondroitin sulphate proteoglycan NG2. Numbers interspersed into gap-containing columns indicate a short stretch of residues of the original sequence that has been omitted here to avoid inflating the alignment. SWISS-PROT (Bairoch & Apweiler, 1997; P or Q as first letter) and PIR accession numbers for sequences in order of the alignment are: A33136, P20908, P12107, P39059, Q07052, P39061, I55398, JP0076, P35440, Q06441, P06205, P02741, P26022, P47970, A38069, A39039, A40228, A48089, B53580, A55347, P31696, B36665, P25391, Q00657 and S28399.

Figure 4 shows an alignment of both consensus strings, TspN (139 residues) and LamG (142 residues), against the pentraxin module CRP-1.1 (146 residues) for which a structural model is available from the PDB data base(code: 1lim, Bernstein *et al.*, 1977). Only 25 residues are identical in the three sequences of the alignment, but 32 further positions have very similar amino acid residues (together 37% of the positions). TspN shares 37 positions with LamG and 30 with CRP-1.1, while the latter two are more closely related, sharing 46 positions (again cf. the scheme in Figure 1). Regions of particular conformity between the three sequences coincide with the beta-strands of the 1lim structure. All these findings together corroborate the conclusion that the three sequences form a common module.

Figure 5 shows an alignment of a representative sample of all three families displaying the common block as well as subtle differences in the non-conserved regions. The difference in known or proposed disulfide bridges is striking, while there are 9 to 11 conforming beta-strands. To confirm these results we submitted several sequences of the three families to the new PSI-BLAST (Altschul *et al.*, 1997). Upon convergence only LamG and TspN-family members were reported with statistical significance.

We also offered the LamG consensus to the fold recognition server (Fischer & Eisenberg, 1996). *L. polyphemus* C-reactive protein (PDB code 1 Lim a; Bernstein *et al.*, 1977) was reported with a *Z*-score of 9.05 (the reliability threshold was stated as a *Z*-score of 4.8 ± 1). The TspN consensus yielded no reliable result. We subjected both consensus strings to TOPITS (Rost, 1995), another prediction-based threading method. Although the *Z*-scores were low (2.1, 2.3), the serum amyloid P component PDB code 1sac) was reported in both cases as most similar fold. Finally, we note here that Moradi-Ameli *et al.* (1994) predicted a sandwich of anti-parallel β strands for the TspN module of the respective collagens.

Our results strongly suggest similarity of TspN and LamG modules to each other and to the *L. polyphemus* C-reactive proteins and other pentraxins. In addition to structurally conserved features shared by all pentraxins, a subfamily consisting of the longer pentraxins and *Limulus* CRPs share a conserved loop region preceding strand h (LADGKQW, position 71 in Figure 5) that is not shared with human Sap or CRP. As TspN and LamG contain a similar region, subfunctions might be shared.

In the course of our analysis, we were able to detect new TspN modules in several proteins. It could be identified in all known Tsps, although Tsps 3 and 4 have been proposed to be distinct in their terminus (e.g. see Lawler *et al.*, 1993; Bornstein & Sage, 1994). This finding is consistent with the fact that all four Tsps bind heparin with their N-terminal region. Furthermore, the modular nel protein and its relatives (Matsushashi *et al.*, 1995; Watanabe *et al.*, 1996) possess a TspN domain.

Despite the vast literature on binding functions we are not able to conclude a common function for all superfamily members. However, some functions are shared at least by members of all three families: thus, the binding of heparin was reported for several LamG modules (e.g. see Yurchenko *et al.*, 1993), for all TspNs in thrombospondins (Bornstein & Sage, 1994), and for human SAP (Li *et al.*, 1994). At the molecular level, the alignment of representative superfamily members (Figure 5) reveals two exposed, hitherto unrecognised conserved positions, both predominantly occupied by aspartate residues. One (position 92 in Figure 5) occurs in all subfamilies and might represent a common functional site, whereas aspartate 72 (Figure 5) is present only in family members that do not contain the signature of the described Ca^{2+}-binding sites of ''short'' pentraxins (Shrive *et al.*, 1996; Emsley *et al.*, 1994). The conservation pattern of aspartate 72 (Figure 5) suggests two possible scenarios: that either the Ca^{2+}-binding of the short pentraxins is retained, but the functional conservation is due to sequence conservation at another site, or that there is no functional conservation and the sequence conservation differing from the ''short'' pentraxins serves another functionality.

We conclude that a number of extracellular module proteins, structural as well as bridge-forming, are united, in spite of all their diversity, by a family relationship. It will be interesting to see whether convergent or divergent evolution was the driving force.

Acknowledgments

We thank the unknown referees for fruitful criticism. We point out that, during revision, we replaced the results obtained with BLAST (Altschul *et al.*, 1990) by use of the new Gapped BLAST (Altschul *et al.*, 1997), which was not available at the time of first submission of our manuscript.

References

Adams, J. & Lawler, J. (1993). The thrombospondin family. *Curr. Biol.* **3**, 188–190.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST, a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.

Bairoch, A. & Apweiler, R. (1997). The SWISS-PROT protein sequence database and its supplement TrEMBL. *Nucl. Acids Res.* **25**, 31–36.

Bairoch, A., Bucher, P. & Hofmann, K. (1997). The PROSITE database, its status in 1997. *Nucl. Acids Res.* **25**, 217–221.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., 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.

Bocchinfuso, W. P. & Hammond, G. L. (1994). Steroid-binding and dimerization domains of human sex hormone-binding globulin partially overlap: steroids and Ca^{2+} stabilize dimer formation. *Biochemistry*, **33**, 10622–10629.

Bork, P. (1992). The modular architecture of vertebrate collagens. *FEBS Letters*, **307**, 49–54.

Bork, P. & Bairoch, A. (1995). Extracellular protein modules. *Trends Biochem. Sci.* **22**, C02 (Supplement).

Bork, P., Downing, A. K., Kieffer, B. & Campbell, I. (1996). Structure and distribution of modules in extracellular proteins. *Quart. Rev. Biophys.* **29**, 119–167.

Bornstein, P. (1995). Diversity of function is inherent in matricellular proteins: an appraisal of thrombospondin 1. *J. Cell. Biol.* **130**, 503–506.

Bornstein, P. & Sage, H. C. (1994). Thrombospondins. *Methods Enzymol.* **245**, 62–85.

Brancaccio, A., Schulthess, T., Gesemann., M. & Engel, J. (1995). Electron microscopic evidence for a mucin like region in chick muscle alpha dystroglycan. *FEBS Letters*, **368**, 139–142.

Delwel, G. O. & Sonnenberg, A. (1996). Laminin isoforms and their integrin receptors. In *Adhesion*

Receptors as Therapeutic Targets (Horton, M. A., ed.), pp. 9–36, CRC Press, Boca Raton.

Deutzmann, R., Huber, H., Schmelz, K. A., Oberbäumer, I. & Hatl, L. (1988). Structural study of long arm fragments of laminin. Evidence for repetitive C-terminal sequences in the A-chain, not present in the B-chains. *Eur. J. Biochem.* **177**, 35–45.

Doolittle, R. F. (1995). The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64**, 287–314.

Emsley, J., White, H. E., O'Hara, B. P., Oliva, G., Srinivasan, N., Tickle, I. J., Blundell, T. L., Pepys, M. B. & Wood, S. P. (1994). Structure of pentameric human serum amyloid P component. *Nature,* **367**, 338–345.

Fischer, D. & Eisenberg, D. (1996). Fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947–955.

Gewurz, H., Zhang, X. H. & Lint, T. F. (1995). Structure and function of the pentraxin. *Curr. Opin. Immunol.* **7**, 54–64.

Goodman, A. R., Cardozo, T., Abagyan, R., Altmeyer, A., Wisniewski, H. G. & Vilcek, J. (1996). Long pentraxins: an emerging group of proteins with diverse functions. *Cytokine Growth Factor Rev.* **2**, 191–202.

Joseph, D. R. & Baker, M. E. (1992). Sex hormone-binding globulin, androgen-binding protein, and vitamin K-dependent S are homologous to laminin A, merosin and *Drosophila* crumbs protein. *FASEB J.* **6**, 2477–2481.

Kohda, D., Morton, C. J., Parkar, A. A., Hatanaka, H., Inagaki, F. M., Campbell, I. D. & Day, A. J. (1996). Solution structure of the link module: a hyaluronan-binding domain involved in extracellular matrix stability and cell migration. *Cell,* **86**, 767–775.

Lawler, J., Duquette, M., Urry, L., McHenry, K. & Smith, T. F. (1993). The evolution of the thrombospondin gene family. *J. Mol. Evol.* **36**, 509–516.

Li, X. A., Hatanaka, K., Guo, L., Kitamura, Y. & Yamamoto, A. (1994). Binding of serum amyloid P component to heparin in human serum. *Biochim. Biophys. Acta,* **1201**, 143–148.

Manfioletti, G., Brancolini, C., Avanzi, G. & Schnieder, C. (1993). The protein encoded by a growth-arrest specific gene (gas6) is a new member of the vitamin K dependent proteins related to protein S, a negative regulator of the blood coagulation cascade. *Mol. Cell. Biol.* **13**, 4976–4985.

Mark, M. R., Chen, J., Hammonds, G. R., Sadick, M. & Godowski, P. J. (1996). Characterization of Gas6, a member of the superfamily of G domain-containing proteins, as a ligand for Rse and Axl. *J. Biol. Chem.* **271**, 9785–9789.

Matsuhashi, S., Noji, S., Koyama, E., Myokai, F., Ohuchi, H., Tanguchi, S. & Hori, K. (203). New gene nel, encoding a M(r) 93 K protein with EGF-like repeats is strongly expressed in neural tissues of early stage chick embryos. *Dev. Dynam,* **203**, 212–222.

Mayne, R. & Brewton, R. G. (1993). New members of the collagen superfamily. *Curr. Opin. Cell Biol.* **5**, 883–890.

Mercurio, A. M. (1995). Achieving specificity through cooperation. *Trends Cell Biol.* **5**, 419–423.

Moradi-Ameli, M., Deleage, G., Geourjon, C. & van der Rest, M. (1994). Common topology within a non-collagenous domain of several collagen types. *Matrix Biol.* **14**, 233–239.

Osmand, A. P., Friedenson, B., Gewurz, H., Painter, R. H., Hofmann, T. & Shelton, E. (1977). Characterization of C-reactive protein and the complement subcomponent Clt as homologous proteins displaying cyclic pentameric symmetry (Pentraxins). *Proc. Natl Acad. Sci. USA,* **74**, 739–743.

Patthy, L. (1991). Laminin A-related domains in crb protein of *Drosophila* and their possible role in epithelial polarization. *FEBS Letters,* **289**, 99–101.

Patthy, L. (1992). A family of laminin-related proteins controlling ectodermal differentiation in *Drosophila*. *FEBS Letters,* **298**, 182–184.

Patthy, L. (1996). Exon shuffling and other ways of module exchange. *Matrix Biol.* **15**, 301–310.

Rost, B. (1995). TOPITS: threading one-dimensional predictions into three-dimensional structures. *The Third International Conference on Intelligent Systems for Molecular Biology (ISMB)* (Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. & Wodak, S., eds), pp. 314–321, AAAI Press, Cambridge, U.K.

Rothberg, J. M. & Artavanis-Tsakonas, S. (1992). Modularity of the Slit protein. *J. Mol. Biol.* **227**, 367–370.

Shrive, A. K., Cheetham, G. M. T., Holden, D., Myles, D. A. A., Turnell, W. G., Volanakis, J. E., Pepys, M. B., Bloomer, A. C. & Greenhough, T. J. (1996). Three dimensional structure of human C-reactive protein. *Nature Struct. Biol.* **3**, 346–353.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **145**, 195–197.

Srinivasan, N., White, H. E., Emsley, J., Wood, P., Pepys, M. B. & Blundell, T. L. (1994). Comparative analysis of pentraxins: implications for protomer assembly and ligand binding. *Structure,* **2**, 1017–1027.

Ushkaryov, Y. A., Petrenko, A. G., Geppert, M. & Sudhof, T. C. (1992). Neurexins: synaptic cell surface proteins related to the alpha-latrotoxin receptor and laminin. *Science,* **257**, 50–56.

Watanabe, K. T., Katagiri, T., Suzuki, M., Shimizu, F., Fujiwara, T., Kanemoto, N., Nakamura, Y., Hirai, Y., Maekawa, H. & Takahashi, E.-I. (1996). Cloning and characterization of two novel human cDNAs (Nell1 and Nell2) encoding proteins with six egf-like repeats. *Genomics,* **38**, 273–276.

Yurchenko, P. D., Sung, U., Ward, M. D., Yamada, Y. & O'Rear, J. J. (1993). Recombinant laminin G domain mediates myoblast adhesion and heparin binding. *J. Biol Chem.* **268**, 8356–8365.

Edited by G. Von Heijne

(Received 21 July 1997; received in revised form 21 October 1997; accepted 3 November 1997)