JMB



COMMUNICATION

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

Georg Beckmann^{1*}, Jens Hanke¹, Peer Bork^{1,2} and Jens G. Reich¹

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

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

© 1998 Academic Press Limited

*Corresponding author

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 threedimensional (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, aminoterminal thrombospondin; CRP, C-reactive protein; SAP, serum amyloid protein.

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 highestscoring 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



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.

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.

We have developed a software tool called IDSECS (for Iterative Database Search with Evolving Consensus Strings) 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 nonredundant 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.

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



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.

three families, again accumulating several nonmembers 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 superfam-TspN_CONS, LamG_CONS, ilv. 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.

	11	21	31	41	51	61	71
		T I	1				
TSON	PEEFSLLATFR	LAPKSSGTI	FALYOAT	DGVPOFGVVLI	GR-SKILLFYL	KGGRGKTOVVP	FFRI.P-TADCOWN
PARPD	PKDESLLTAVE	ARPGLOAPI	LTLYSA	OGVROLGLEL	GR PVRFLVED-	OTGRDODDAOR	VERGLELADOWNE
COLSAID	DEDESTLTTUE	VARGEOVAL	VCTVNP	OGTOOTOL PL	OR COURT VED	UTORDODROVE	I PROINT ODOKWI
COLLINIA	DEDECTI PEUK	DERGIOGRE	TOTWNP		GR OPUPLED	HIGKPGPEDIP	LFRGINLSDGRWA
COLIENT	PEDISIDEIVK	PREGIQSEL	LSIINE	HGTQQIGVEV	GRSPVFLFED-	HIGKPAPEDYP	LFRIVNIADGKWH
COLISAIN	FRDFAIRLVVK	PSSTRGGVI	FAITDAF(QKVIYLGLRLSGVEL	GH-QRIILYYTE-	PGSHVSQEAP	PAFSVP-VMTHRWN
COL16A1h	PEEFALVLTLL	LKKHTHQK1	WYLFQVTDA-1	NGYPQISLEVN	ISQERSLELRAQG-	QDGDFVSCIFP	VPQLFDLRWH
COL18A1h	FRDFSLLFHVR	PATEAAGVI	FAITDAA(QVVVSLGVKLSEVRI	GQ-QNISLLYTE-	PGASQTQTGA	SFRLP-AFVGQWT
TSP3m	AGDIYLLSTFR	LPPKQGGVI	FGLYSRQI	DNTRWLEASVV	GKINKVLVRYQR-	EDGKVHAVNLQ	QAGLADGRTH
NELC	KYEFTILVTLK	QAHLNSGVI	FSIHH	LDHRYLELESS	GHRNEIRLHYRT-	GSHRSHTEVFP	YILADDKWH
TSP1c	NEGFILSATLE	ODROSRGTI	LALEGPG	ISEROFETISN	GRANTLDLIYWV-	DGFONVISLE	DVD LADSOWK
TSP4v	MNEUVULSTER	LODKSTWT	PGLVSTS	NOPPERTU	OPI NYA CI DVI D	CDOVI NOUPPN	VID
1 GNH	000000000	a 333 A	AAAAAA	00000 0	£66666	aaaaaa	KUD IADORQI
11.75	444446666666	E 6666	666666	222222	5 ILLILL	999999	11
CDD111	1444440000		000000	///////////////////////////////////////	77 8888888	999999	aa
CRPIII	LOBITICIWER	V-NRLKGTL	HMFSYATAI	KKDNELLTLII	BC-GDEPENAH	GAPQLKVQC-P	NKIHIGKWH
CRPh	LKAFTVCLHFY	TELSSTRG	SIFSYATK	RQDNEILIFWS	SKD-IGYSFTV	GGSEILFEV-P	EVTVAP-V
PTX3h	LESFSACIWVK	A-TDVLNKI	TLFSYGTK1	RNPYEIQLYLS	SYQSIVFVV	GGEENKLVA-E	AMVSLGRWT
APEXg	LYSFTICLWLR	S-SASPGIG	TPFSYAVP(GQANEIVLIEWG	NNPIELLI	NDKVAQLP	LFVSDGKWH
LamG	VPEIDLSLQFR	T-TQPNGLL	LYAGGKI	NGNDFLALELIE	GRLVFLFDL-	GNGRLKVQS-P	NKLNDGKWH
HSPG2h	EVPETIELEVR	T-STASGLI	LWQGVEVGE2	GKDFISLGLQE	GHLVFRYQL-	GSGEARLVS-E	DPINDGEWH
ABP1r	ISKPSSSFEFR	T-WDPEGVI	FYGDTN	FEDDWFMLGLR E	GOLEIOLH	NLWARLTVGF	GPRLNDGRWH
NEU1Ar	KKTGSTSEDER	T-TEPNGL	LESHGKPRH9	KUDEEATEML	GH LVLLLDM-	GSGTTETEALO	KK VNDGRWY
CASEm	DTDI.LAFFDFP	T-FDDFOUT	FFACCP	CDOTWINI OF D 2	OR TRIOLDYN		TTNU(MWO)
NETTOPh	TDODDI AUGRO	T TUKDOT	URIDEAD (DOTAL VIOLAN	OR TOWNENT	COUNT CIRDED	
NE03BD	TRODREAVER 3	T-TVKDGIL	VKIDSAF(SUGDFLQLHIEQ	GRIGVVFNI-	GIVDISIKEER	TPVNDGKIH
LAMASI	KPRSQFAVDMQ	T-TSSRGLV	FHTG	TKNSFMALYLSF	GRLVFALGT-	DGKKLRIKS-K	EKCNDGKWH
AGRNC	YHTVRIAMEFR	A-TELSGLL	LYNGQNI	RGKDFISLALVG	GFVELRFNTG	SGTGVITSKVR-	VEPGKWH
SLIT2d	RPEANVTIVFS:	S-AEQNGIL	MYD(GODAHLAVELFN	GRIRVSYDV-	GNHPVSTMYSF	EMVADGKYH
FATd	VTTNDISIVFA'	T-TKPNSLL	LYNYGMQSG-(GRSDFLAIELVE	DRAYFSSGGAI	RTAISTVIAGRN	LADGGWH
LAMA1h	RKKLSVELSIR'	T-FASSGLI	YYMAHQN	NQADYAVLQLHG	GR LHFMFDL -	GKGRTKVSH-P	ALLSDGKWH
NG2r	LTRVDLLLQFS'	T-SQPEALL	LLAA(GOTDHLLLQLQS	GHLQVRLAL-	GQNELSLQT-P	ADTVLSDSTTH
LAMAd	RRHHDIGISFR'	T-ERPNGLL	IYAGSK(RDDFIAVYLLD	GRVTYEIRV-	GAOLOAKITT-E	ABLNDGTWH
	81	91	101	111	121		131 139
	81 	91 	101	111	121 		131 139
TspN	81 RLALSVSGS-SV	91 VTLYV <mark>D</mark> CNK	101 IDRRPLE	111 PRPFQGLIDVD G TIV	121 LGTRANKGQ	PFQ	131 139 GELQQLKIVCD
TspN PARPb	81 RLALSVSGS-ST RVAVAVKGQ-ST	91 VTLYVDCNK VTLIIDCKK	101 IDRRPLE	111 PRPFQGLIDVD G TIV PRSARPVLDTR G VII	121 LGTRANKGQ FGARILDEE	PFQ	131 139 GELQQLKIVCD GDIQELSIIPG
TspN PARPb COL5A1h	81 RLALSVSGS-ST RVAVAVKGQ-ST RIALSVHKK-NT	91 VTLYVDCNK VTLIIDCKK VTLILDCKK	101 IDRRPLE RVTRPLE	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVII DRSDHPMIDINGIIV	121 LGTRANKGQ FGARILDEE FGTRILDEE	PFQ VFE VFE	131 139 GELQQLKIVCD GDIQELSIIPG GDIQQLLFVSD
TspN PARPb COL5A1h COL11A1h	81 RLALSVSGS-ST RVAVAVKGQ-ST RIALSVHKK-NT RVAISVEKK-T	91 VTLYVDCNK VTLILDCKK VTLILDCKK VTMIVDCKK	101 IDRRPLE RVTRPLE KTTKPLE	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVII DRSDHPMIDINGIIV DRSERAIVDTNGITV	121 IGTRANKGQ FGARILDEE FGTRILDEE	PFQ VFE VFE VFE	131 139 GELQQLKIVCD GDIQELSIIPG GDIQQLLFVSD GDIQQFLITGD
TspN PARPb COL5A1h COL11A1h COL15A1b	81 RLALSVSGS-ST RVAVAVKGQ-ST RIALSVHKK-TT RVAISVEKK-TT RFAMIVOGE-ST	91 VTLYVDCNK VTLILDCKK VTLILDCKK VTMIVDCKK VTLLVNCEE	101 IDRRPLF KTTRPLI KTTKPLI KTTKPLI	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVII DRSDRAIVDINGIIV DRSERAIVDTNGITV SSOALAFESSAGIF	121 IGTRANKGQ FGRRILDEE FGTRILDEE MGNAGATGLE	PFQ 	131 139 GELQQLKIVCD GDIQQLLFVSD GDIQQLLFVSD GDIQQFLITGD GSLQQLTVHPD
TspN PARPb COL5A1h COL11A1h COL15A1h COL15A1h	81 RLALSVSGS-ST RVAVAVKGQ-ST RIALSVHKK-NT RVAISVEKK-TT RFAMIVQGE-ET KIMLSVAGR-V	91 VTLYVDCNK VTLIDCKK VTLIDCKK VTLIVDCKK VTLLVNCEE ASVHVDCSS	101 IDRRPLH KTTKFLI KTTKFLI HSRIPFQI ASSOI	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVII DRSDHPMIDINGIIV RSSQALAFESSAGIF PLGPRRPMRPVGHVG	121 IGTRANKGQ FGARILDEE FGTRILDEE MGNAGATGLE IGLDAECGK	PFQ 	131 139 GELQQLKIVCD GDIQELSIIPG GDIQQLIFVSD GSLQQLTVHPD FDLQQVHIYCD
TSDN PARPb COL5A1h COL13A1h COL15A1h COL16A1h COL16A1h	81 RLALSVSGS-57 RVAVAVKGQ-57 RIALSVHKK-N7 RVAISVEKK-T7 RFAMIVQGE-67 KLMLSVAGR-V1 HEALSUNGG-57	91 VTLYVDCNK VTLIDCKK VTLIDCKK VTLIVNCEE ASVHVDCSS VALVDCEE	101 IDRRPLB RVTRPLB KTTKFLI KTTKPLI ASSQI	111 PRPFQGLDVDGTIV PRSARPVLDTRGVI PRSDHPMIDINGIV PRSERAIVDINGIV RSSQALAFESSAGI LIGPRRPMRPVGHVF HSSQCLENECAGLE	121 IGTRANKGQ FGARILDEE FGTRILDEE MGNAGATGLE IGLDAEQGK IGLDAEQGK		131 139 GELQQLKIVCD GDIQELSIIPG GDIQQLFUTGD GSLQQLTVHPD FDLQQVHIYCD GMISELKVRKT
TspN PARPb COL5Alh COL1Alh COL15Alh COL16Alh COL18Alh	81 RLALSVSGS-S' RIALSVHKK-N' RVAISVEKK-T' RFAMIVQGE-B' KLMLSVAGR-VJ HFALSVDGG-S'	91 VTLYVDCNK VTLILDCKK VTLILDCKK VTLILVNCEE &SVHVDCSS VALYVDCEE	101 IDRRPLI RVTRPLI KTTKFLI KTTKPLI HSRIPFQR FQRVPFAR	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVII DRSDHPMIDINGIIV DRSBRAIVDTNGIFV SSQALFESSAGIF PLGPRPMRPVGHVF PASQCLELERCAGLE DIA DE DDB PUGCH	121 IGTRANKGQ FGRRILDEE FGRRILDEE FGRRILDEE IGLDAEQGK VGQAGTADPD VGCQAGTADPD	- PFQ - VFE - VFE - VFE - VFE - VFE - PFT - PFS - VFE - PFS - VFE - PFS - PFS	131 139 GELQQLKIVCD GDIQELSIIPG GDIQQLFVSD GDIQQFLITGD GSLQQLTVHPD FDLQQVHIYCD GMISELKVRKT GPVESMKTILG
TSDN PARPb COL15A1h COL15A1h COL15A1h COL16A1h COL18A1h TSP3m	81 RLALSVSGS-ST RVAVAVKGQ-ST RIALSVHKK-NT RVAISVEKK-T RFAMIVQGE-ST KLMLSVAGR-VJ HFALSVDGG-ST TALLRLRGP53	91 VTLYVDCNK VTLILDCKK VTLILDCKK VTLLVNCEE ASVHVDCSS VALYVDCEE LQLYVDCKL	101 IDRRPLH RVTRPLI KTTKFLI HSR1PFQI ASSQI FQRVPFAH GDQHAGLI GDQHAGLI	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVI PRSBRAIVDTRGIV SSSDALPESSAGI PLGPRRPMRPVGHVF PALAPIPPAEVGGUF PALAPIPPAEVGGUF	121 GTRANKGQ FGTRILDEE FGTRILDEE MGNAGATGLE UGLABAEGK VGQAGTADPD IIRTGQKAYL IIRTGQKAYL	- PFQ - VFE - VFE - VFE - VFE - PFS - VFE - PFS - VFE - VFE - PFS - VFE - PFS - VFE - PFS - VFE - PFS - VFE - PFS - PFS	131 139 GELQQLKIVCD GDIQELSIIPG GDIQQLLFVSD GDIQQLTVHPD FDLQQVHIYCD GMISELKVRKT GFVESMKIILG GFVESMKILG
TspN PARPb COL5A1h COL15A1h COL15A1h COL16A1h COL18A1h TSP3m NELg	81 RLALSVSGS-S' RVAVAVKGQ-S' RIALSVHKK-TT RVAISVEKK-TT RFAMIVQGE-B' KLMLSVAGR-V HFALSVDGG-S' TALLRRGP50 RLSLAISAS-H	91 VTLYVDCNK VTLILDCKK VTLIVDCKK VTLIVNCEE ASVHVDCSS VALYVDCEE LQLYVDCKL LILHVDCNK	101 	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVII PRSBRAIVDTNGITV DRSBRAIVDTNGITV USSQALAFESSAGIF PLGPRRPMRPVGHVF VASQGLELERGAGLF PALAPIPABVSGLE VEKPFMDLPVGTTFW	121 GTRANKGQ FGARILDEE FGTRILDEE GGTRILDEE UGLDAEGAK UGQAGTADPD ILGQARAPL LLGQRNNAHG LLGQRNNAHG		131 139 GELQQLKIVCD GDIQELSIIPG GDIQQLLFVSD GSLQQLTVHPD FDLQQVHIYCD GMISELKVRKT GFVESMKIILG GIMQDVQLLVM
TspN PARPb COL5Alh COL15Alh COL15Alh COL16Alh COL18Alh TSP3m NELg TSP1c	81 RLALSVSGS-S1 RVAVAVKGQ-S7 RIALSVHKK-NT RVAISVEKK-TT RFAMIVQGE-E KLMLSVAGG-S7 TALLRLRGP5G TALLRLRGP5G RLSLAISAS-H NLTVQVTGE-NT	91 VTLYUDCNK VTLILDCKK VTLLLDCKK VTLLLOCKK VTLLVNCEE &SVHUDCSS VALYUDCEL LQLYUDCKL LLLHVDCNK YNLYVGCDL	101 	111 PRPFQGLIDVDGTIV PRSRAPVLDTRGVI PRSERATVDTNGIV SSQALAFESSAGIF VLGPRRFMRFVGHVF RASQGLELERGAGLE PALAPIPPAEVSGLE ERFPMLPVGTTF# 2000	121 LGTRANKGQ FGRRILDEE FGRRILDEE MGNAGATGLE VGQAGTADPD		131 139 GELQQLKIVCD GDIQQLLSIPG GDIQQLLTGD GSLQQLTVHPD GMISELKVRKT GFVESMKIILG GINQDVQLVM GLUQNIHLIFD GSUDEU UVC
TSDN PARPb COLIAIH COLIAIH COLISAIH COLISAIH TSP3m NELg TSP1c TSP4x	81 RLALSVSGS-57 RVAVAVKGQ-57 RIALSVHKK-TY RFAMIVQGE-57 KLMLSVAGR-47 TALLRLRGP5G TALLRLRGP5G TALLRLSAS-H NLTYQVTGE-NT ALLLHSGS5F/	91 VTLYVDCNK VTLILDCKK VTLLUDCKK VTLLVNCEE ASVHVDCSS VALYVDCEL LQLYVDCKL LLIHVDCNK VNLYVGCDL AKLYIDCNE	101 	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVII DRSDFMIDINGITV SSSBAIVDRGITV SSSBAIVDRGITV SSSALAFESSAGIF PLGPRRPMRPVGHVF PALAPIPPAEVSGLE JEKPFMDLPVCTTFW PRPLSGIRLNTGSV	121 LGTRANKGQ FGRRILDEE FGTRILDEE MGNAGATGLE LGLDAEQGK IRTGQKAYL LGQRNNAHG LGQRNNAHG LGRENNAHG		131 139 GELQQLKIVCD GDIQELSIIPG GDIQQLFVSD GSLQQLTVHPD FDLQQVHIYCD GMSSELKVRKT GFVESMKIILG GILQNIHIFD DSXDDELKVMG
TspN PARPb COL5Alh COL11Alh COL15Alh COL15Alh COL15Alh TSP3m NELg TSP1c TSP4x IGNH	81 RLALSVSGS-SJ RVAVAYKG-SJ RIALSVHKK-N RVAISVEKK-T RFAMIVQGE-B KLMLSVAGR-V HFALSVDGG-SJ TALLRRGP5G RLSLAISAS-H NLTVQVTGE-N ALLLHLSGL5FJ ALLLHLSGL5FJ	91 VTLYDCNK VTLILDCKK VTLILDCKK VTLILDCKK VTLUNCES ASVHVDCSE LQLYVDCKL LLLHVDCNK VLLYDCES LQLYVDCKL LLLHDCNK YNLYVGCDL AKLYLDCNF iiiii jj	101 	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVII PRSSHPMIDINGIIV PRSSRAIVDINGIV SSQALAFESSAGIF VLGPRRFMRFVGHVF RASQGLELERCAGLE PALAFIPABUSGLE PREKPFMDLPVGTTF# PRPLSGIRLNTGSVF kkk	121 LGTRANKGQ FGRRILDEE FGTRILDEE MGNAGATGLE UGLDAEGK VGQAGTADPD IRTGQKAYL LGQRNNAHG VAKOSIREN		131 139 GELQQLKIVCD GDIQQELSIIPG GDIQQFLITGD GDLQQLFVYDD GDLQQLTVHPD GMSDELKVRKT GFVESMKILG GINQDVQLLVM GLUQNIHLIFD DSMDELK.VMG 1111111111
TSDN PARPb COL5A1h COL15A1h COL15A1h COL16A1h COL16A1h TSP3m NELg TSP1c TSP4x 1GNH 1LIM	81 RLALSVSGS-S' RVAVXKQC-S' RIALSVHKK-T' RYAISVEKK-T' RFANIVQGE-S' TALLR.RGP5G RLSLAISAS-H MLTVQVTGE-N' ALLLH.SGL5F, hhhhhhh i:	91 VTLYVDCNK VTLILDCKK VTLILDCKK VTLILDCKK VTLUNCEE a\$VHVDCSS VALYVDCEE LQLYVDCKL LLHVDCNK YNLYVGCDL aKLYLDCNF iiiiij jj	101 IDRPLI KVTRPLI KTTKPLI HSRIPFQI HSRIPFQI GDQHAGLI IYERV IYERV IYERV TGVVEDLI jj :	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVI PRSBRAIVDTRGIV VSSERAIVDTGITV SSQLAFESSAGIE PLGPRRPMRPVGHVF PALAPIPPAEVGGUF JEKPFMDLPVGTFW PRPLSGIRLNTGSVW kkk dddd	121 LGTRANKGQ FGRRILDEE FGRRILDEE KGNQATGLE LGLDAEQGK UGQATADPD LIRTQQKAYL LGQRNNAHG		131 139 GELQQLKIVCD GDLQELSIFG GDLQQLLFVSD GDLQQLTVHDD FDLQQVHIYCD GMXSELKVRKT GFVESMKIILG GILQNIHLIFD DSMDELKLVMG 111111111
TepN PARPb COLSAIh COLISAIh COLISAIh COLISAIh COLISAIM NELg TSPic TSP4x IGNH ILM CRP111	81 / RLALSVSGS-SJ RVAVAYKG-SJ RIALSVHKK-N RVAISVEKK-T RFANIVGE-EJ KLMLSVGG-SJ TALLRLRGP5G TALLRLRGP5G NLSLAISAS-H NLTVQVTGE-N ALLLHLSGL5F, hhhhhhh i: aaaaaaaa20	91 // VTLYUDCNKK VTLILDCKK VTLILDCKK VTLILDCKK VTLIVDCKL VTLIVDCKL VTLIVDCKL VTLIVDCKL VILIVGCL AKLYIDCNF iiiii jj bbbbbb cc ATIAVDGFH	101 	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVII PRSERAIVDTRGITV INSERAIVDTRGITV INSERAIVDTRGITV INSERAIVDTRGITV INSERAIVDTGITV RALAPIPPAEVSGLE JEKPFMDLPVGTTVF ISK PFYEQLKAENSRMY PRPLSGIRLNTGSVF kkk dddd SIAVGRTLSQGGLVV	121 LGTRANKGQ FGRRILDEE FGTRILDEE FGTRILDEE VGQAGTADPD UGQDAGTADPD LGQRNAHG		131 139 GELQQLKIVCD GDIQQELSIIPG GDIQQFLTGD GDLQQLFVPD GSLQQLTVHPD FDLQQVHIYCD GMSSELKVRKT GIVSSMKILG GIVSSMKILG GIVQDVQLLVM GLLQLNHHIFD DSMDELKVMG 1111111111
TSPN PARPb COLSAIh COLIAIh COLIAIh COLIAIh TSPic TSPic TSPic TSP4x IGNH ILTM CRP111 CRP1	81 RLALSVSGS-SJ RIALSVHK-TN RYALSVEK-TN RYALSVEK-TN RFANIVQGS-SJ TALLRRGP5G HLSLAISAS-H NLTVQVTGS-N ALLLHSGLSF: aaaaaaaaa2bJ HVCHTNSSW2E HUCTSWESA2T	91 VTLYVDCNK VTLIDCKK VTLIDCKK VTLUDCKK VTLVDCK ASVHVDCSS ALYVDCH LILHVDCNK VNLYVGCDL AKLYIDCNF IIIII DbDbDbD cc ATIAVDGFH WFWVDGK	101 IDRRPLH KVTRPLI KTTKPLI KTTKPLI FQRVPFAI FQRVPFAI FQRVPFAI IDSFILEH TGVVEDLI jj cc CKGNATK	111 PRPFQGLIDVDGTIV PRSRAPVLDTRGVI PRSBATVDTNGIV SSSDLAFESSAGIF PLGPRFPMRPVGHVF PLGPRFPMRPVGHVF PLSGFLERGAGLF PRPLSGIRLNTGSVF kkk dddd GIAVGRTLSQGGL LKKGYTVGZEASII	121 LGTRANKGQ FGRRILDEE FGRRILDEE MGNAGATGLE VGQAGTADPD IRTGQKAYL VAKGSIREN		131 139 GELQQLKIVCD GDLQLLFVSD GDLQQLLFVSD GDLQQLTVHPD GSLQQLTVHPD GSLQQLTVHPD GMSELKVRKT GLVQNIHLIFD DSMDELKLVMG GLLQNIHLIFD GDLGNVMNMDF
TapN PARPb COLSA1h COLISA1h COLISA1h COLISA1h TSP3m NELg TSP4x 10NH 1LIM CRP111 CRP111 CRP11	81 RLALSVSGS-S' RVAVAVKGQ-S' RVAISVEKK-T' RPAMIVQGE-E' KLMLSVGR-U' TALLRLRGP5G' RLSLATSAS-H HALLKGESF/ hhhhhh 1 ALLLHSGLSF/ hhhhhhh 1 HUCTYNSE2L' HUCTYNSE2L'	91 / VTLYVDCNK VTLILDCKK VTLLLDCKK VTLLLDCKK VTLLUNCEE SVHVDCSS ALYVDCEL ALYVDCEL ALYVDCKL ALYUDCKL ALYIDCNF IIIII JJ Dobbbbb c ATIAVDGPH VEFWVDGKE TSLWVNGEL	101 IDRPLI RVTRPLI KTTKFLI HSRIPFQI SSQI FQRVPFAI GDQHAGLI IVERVU IVERVU IVERVU TYERVE 	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVII PRSERAIVDTRGITV PRSERAIVDTRGITV RSSERAIVDTRGITV RSSERAIVDTRGITV RSSERAIVDTGITV PALAPIPPAEVSGLE PALAPIPPAEVSGLE PREVENSION PRPLSGIRLNTGSVE kkk dddd SIAVGRTLSQGGLVT LKKGYTVGAEGSLI	121 LGTRANKGQ FGARILDEE FGTRILDEE MGNAGATGLE LGLDAEQ3K LGLDAEQ3KAYL LGQRNNAHG LGQRNNAHG LGQRNNAHG k ldd LGQEQDSFGG LGQEQDSFGG LGQEQDSFGG		131 139 GELQQLKIVCD GDLQLLFVSD GDLQQLFVSD GDLQQLTTGD GSLQQLTUHPD FDLQQVHIYCD GMSSELKVRKT GFVBSMKIILG GLQNIHLIFD DSMDELKLVMG 11111111111 GELSELNLWMT GDLGNVNMMDF GRITGFNIWDS
TspN PARPb COL5Alh COL1Alh COL15Alh COL16Alh TSP3m TSP1c TSP4x 1GNH 1LTM CRP111 CRP111 CRP1	81 RLALSVSGS-SJ RVAVAYGQ-SJ RIALSVHKK-N RYALSVEK-T RFAMIVQGE-SJ TALLRRGP5G KLMLSVAGR-V, HFALSVDGG-SJ TALLRGF5G HLSLATSAS-H NLTVQVTGE-N ALLLHSGL5T HLCTNSS22T HLCTTNSS22T HLCTTNSS22T	91 // VTLYVDCNK VTLIIDCKK VTLIIDCKK VTLUDCKK VTLUDCKK VTLVDCKK SVHVDCSS SUJVYDCKI SUJVYDC	101 IDRRPLI KTTKPLI KTTKPLI KTTKPLI FQRVPFAI FQRVPFAI IDSFILEI IDSFILEI IDSFILEI CKGNATG CKGNATG 	111 PRPFQGLIDVDGTIV PRSRAPVLDTRGVI PRSSRATVDTNGIV PRSSRATVDTNGIV SSQALAFESSAGIF PALAPIPAEVSGLE UGPRRPMRPVGHVF PRPLSQRLAENSRW PRPLSQRLAENSRW PRPLSQRLATUSQCGLV LKKGYTUGABASII MATGHIVPESGILQ LALPHPIKSGGULU	121 LGTRANKGQ FGRRILDEE FGTRILDEE FGTRILDEE VGQAGTADPD UGQAGTADPD URTGQKAYL VGAGTADPD URTLQK k Idd LGQEQDS VGG LGQEQDS VGG LGQEQDS VGG		131 139 GELQQLKIVCD GDIQQELSIIPG GDIQQLFLTGD GDIQQLFUYDD GMISELKVRKT GFVESMKILG GFVESMKILG GINQDVQLLVM GLNQDVLLVMG GLNQDVQLLVMG IIIIIIIIII eeefffffff GELSGENLWMT GGLSUNMWDF GRLTGFNIWDS
TapN PARPb COLSAIh COLIAAIh COLIAAIh COLIAAIh TSP3m NELg TSP4x IGNH ILIM CRP111 CRP111 CRP111 PTX3h APPXg IamC	81 RLALSVSGS-S' RVAVXKGQ-S' RIALSVHKK-T' RYAISVEKK-T' RYAISVEKK-T' RLALSVGG-S' TALLR.RGP5G RLSLAISAS-H HALSVGG-S' ALLLH.SGL5F, hhhhhhh i: NLTVQVTGE-N' ALLH.SGL5F, Hhhhhhh i: UCTTNSS2L' HICTTNSS2L' HICTTNTS22L' HICTTNTT22L HICTTNTT22L	91 VTLYDCNK VTLILDCKK VTLILDCKK VTLILDCKK VTLILDCKK VTLVNCEB ASVHVDCS LDLYVDCKL LLHVDCNK VLYVDCKI IIIII JJ Dobbbb c ATIAVDGFH VEPWVDGKF TSLWVNGEI WEBACDGEK	101 IDRPLI KVTRPLI KTTKPLI HSRIPFQI GDQHAGLI TGVVFAI TGVVFAI TGVVEDLI jj CKONATT CKONATT CKONATT LGTGEN LGTGEN	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVI PRSBRAIVDTRGIV PRSBRAIVDTGITV SSQLAPESSAGIE PLGPRPMPPVGHVF ALAPIPPAEVGGL PFYPQLKAENSRMY PRPLSGIRLNTGSVW ddd dIAVGRTLSQGGLV LKKGYTVGAEASII SMATGHIVPEGGLQ LLAPMHPIKSGGLL LAPMHPIKSGGLL	121 LGTRANKGQ FGRRILDEE FGRRILDEE LGLDAEQGK LGLDAEQGKA LGQRNNAHG LGQRNNAHG LGQRNAHG LGQRDNAHG k LGQRQDSFGG LGQEQDSFGG LGQEQDTVGG LGQEQDTVGG		131 139 GELQQLKIVCD GDLQELSIFG GDLQQLLFVSD GDLQQLLFVSD GDLQQLTVHDD FDLQQVHIYCD GMXSELKVRKT GFVESMKIILG GFVESMKIILG GLLQNIHLIFD DSMDELKLVMG 111111111 GELSELNLWNT GRLTGFNIWDS GRLTGFNIWDS GELSQVINGK
TspN PARPb COL5Alh COL5Alh COL13Alh COL13Alh COL13Alh TSP3m NELg TSP1c TSP4x 10NH 1LTM CRP11 CRP11 CRP11 CRP11 CRP11 CRP14 Lamo	81 RLALSVSGS-ST RVAVAYKG-ST RIALSVHKK-N RVAISVEKK-T RFAMIYGE=B TALLRUP5G LLLAUSUDG-ST ALLLHLSGLSF, ALLLHLSGLSF, HICTSWESA2T HICTSWESA2T HICTSWESA2T HICTSWESA2T VVFTSGR-N	91 / //TLIYDCNK VTLIJDCKK VTLIJDCKK VTLIJDCKK VTLIJDCKK VTLIJDCKK VTLIJDCKK VTLIJDCKK VILIVCKK VILIVCKK VILIVCKK VILIVDCKK VILIVCKK VILIVCK	101 IDRRPLI RVTRPLI KVTKFLI KTKFLI GDQHAGLI GDQHAGLI GDQHAGLI TGVVEDLI jj TGVVEDLI jj CGNATT RVR-KS	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVII PRSSHPMIDTNGITV PRSSHAIVDTNGITV PRSSRAIVDTNGITV PRSSRAIVDTNGITV PALAPIPAEVSGLE PRFVBULPVGTTFW RASGGLELERGAGIE MATGHIVPEGGILO LLKGYTVGAEASII MATGHIVPEGGILO DEGDNUAVMEGGTI	121 LGTRANKGQ FGRRILDEE FGTRILDEE FGTRILDEE VGQAGTADPD VGQAGTADPD VIRTGQKAYL IRTGQKAYL ILGQEQDSVGG LGQEQDSVGG ILGQEQDSVGG ILGQEQDSVGG ILGQEDSKNCC-VGG ILGQENSKNCC-VG ILGQENSKNCC-VG ILGQENSKNCC-VG ILGQENSKNCC-VG ILGQENSKNCC-VG ILGQENSKNCC-VG ILGQENSKNCC-VG ILGQENSKNCC-VG ILGQENSKNCC-VG ILGQENSKNCC-VG ILGQENSKNCC-VG ILGQENSKNCC-VG ILGQENSKNCC-VG ILGQENSKNCC-	PFQ 	131 139 GELQQLKIVCD GDIQQELSIIPG GDIQQFLITGD GDLQQLFVYDD GDLQQLTVHPD FDLQQVHYCD GMSDEQLVW GLVQNIHLIFD DSMDELK.VMG 1111111111 eeefffffff GELSELNUMT GDLGNVNMWDP GRLTGFNIWDS GELSQFNIWDR GCUKNLVHSA
TspN PARPb COL5Alh COL15Alh COL15Alh COL16Alh TSP3m NELg TSP1c TSP4x ISP4 ISP4 CRP111 CRP11 CRP11 CRP11 APEXg LamG HSP02h	81 RLALSVSGS-ST RVAJVKGC-ST RIALSVHKK-TT RFAMIVQGE-ST LLLR.RGP5G HFALSVDGG-ST TALLR.RGP5G HLLR.RGP5G HLLLR.GDSF iaaaaaaaaa2b HUCHTSS242E HLCGTNNSE2L HLCGTNNSE HLCGT	91 // // // // // // // // // /	101 	111 PRPFQGLIDVDGTIV PRSRAPVLDTRGVI PRSSDAPVLDTRGVI PRSSDAPVDTNGIV SSQLAPESSAGIF PLGPRPMRPVGHVF PLGPRLPRAVGGTFW BPFPBQLKAENSRWY PRPLSGIRLNTGSVW LKKGYTVGAEASII SMATGHIVPEGGILG LAPWHPIKSGGVL LAPWHPIKSGGVL SGENVAVNAKGSIF	121 LGTRANKGQ FGRRILDEE FGRRILDEE VGQAGTADPD VGQAGTADPD VIRTGQKAYL VAKGSIREN		131 139 GELQQLKIVCD GDLQELSIFG GDIQQLFVSD GDIQQLFVSD GSLQQLTVHPD FDLQQVHYCD GMSELKVRKT GEVQULVM GLUQNIHLIFD DSMDELKLVMG GULQNIHLIFD GDISNVNMWDF GGLSQFNIWDS GGLSQFNIWDR GCLKDVILNGK GCVCNULVHSA
TspN PARPb COLSAIH COLIAIH COLISAIH COLISAIH COLISAIH TSP3m TSPic TSP4x IGNH 1LTM CRP11 CRP11 CRP11 CRP11 CRP11 HSP62h LamG HSP62h	81 RLALSVSGS-ST RVAVAYKG-ST RIALSVHKK-N RVAISVEKK-T RFANIVGE-ET ILLRUP5G ILLLRGP5G TALLRUP5G-ST ALLLHSGLSF, hhhhhhh i: aaaaaaa2b HICTSVESA2T HLCGTWSS2L HICTSVESA2T TVVFTRSGR-R VTALRGGR-R VTALRGR-R	91 VTLIYDCNK VTLIYDCNK VTLIDCKK VTLIUDCKK VTLIUDCKK VTLIVDCKK ULIVDCKK ULIVDCKK ULIVDCKK ULIVDCKK VLIV VLIVDCKKK VLIVDCKKK VLIVDCKKK VLIVDCKKK VLIVDCKKK VLIVDCKKK VLIVDCKKK VLIVDCKKK VLIVDCKKKK VLIVDCKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKK	101 	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVII PRSBRAIVDTRGITV PRSERAIVDTRGITV RSSERAIVDTRGITV RSSERAIVDTRGITV RSSERAIVDTRGITV PALAPIPPAEVSGLE PALAPIPPAEVSGLE PALAPIPPAEVSGLE PALAPIPPAEVSGLE PREVENTION RV RV RV RV RV RV RV RV RV RV	121 LGTRANKGQ FGRRILDEE FGTRILDEE FGTRILDEE VGQAGTADPD VGQAGTADPD LGQDRNAHG VGQCDVGG LGQDQDSVGG LGQDQDSVGG LGQEQDSVGG LGQEQDTVGG LGQEQDTVGG LGQEQDTVGG LGQLSPSKPKLGG LGQLSPSKPKLGG	PPQ 	131 139 GELQQLKIVCD GDIQQELSIIPG GDIQQFLTGD GDIQQFLTGD GSLQQLTVHPD FDLQQVHYCD GMISELKVRKT GFVESMKILG GIMQDVQLLVM GLLQNIHLIFD DSMDELKLVMG 11111111111 eeefffffff GELSELNLWNT GELSQFNIWDR GCLKDVILNGK GCVKNLVLHSA GCVKNLVLHSA GCVKNLVLHSA
TspN PARPb COL5Alh COL15Alh COL15Alh COL16Alh TSP3m NELg TSP1c TSP4x 1GRH 1LTM CRP111 CRP11 CRP11 CRP11 CRP11 LamG HSPG2h ASP17 NEULAT	81 RLALSVSGS-ST RVAVAVKGC-ST RIALSVHKK-T RPAMIVGB-ST TALLRRGP5G KLMLSVAGR-V HFALSVDGG-ST TALLRRGP5G-N ALLLHSGLSFH HUTVVTTRSA HUCTSWESA2T HLCGTWNSS2LI HUCTSWESA2T HLCGTWNSS2LI HUCTSWESG-N RVTALRGR-R PVELKNRGD-S EVDFQRDGR-S	91 / //TLIYDCNK VTLIYDCKK VTLIYDCKK VTLIVDCK VTLIVDCK SVVDCSL LLHVDCNK VLIVYDCH LLHVDCNK VLIVYOCH VILYVGCH VILYVGCH VILYVGCH VILYVGCH VILYVGCH VILYVGCH VILYVGCH VILVGCH VILVVGC	101 	111 PRPFQGLIDVDGTIV PRSRAPVLDTRGVI PRSSRATVDTNGIV PRSSRATVDTNGIV SSQALAFESSAGIF PALAPIPAEVSGLE UGPRRPMRPVGHVF PRPLSGILLPCSTFW CASQGLELERGAGLE SALAPHOLST MATCHIVPGGIL ULAPHPIX KSGVLI PGNSQLSLPRGLPS PSGPNVAVNAKGSTI PGSSLDHPQLSMRI PGSSELLDLDBELLSMRI PGSSLDHPQLSMRI PGSSELLDLDBELSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPGLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPQLSMRI PGSSLDHPSSLDHPGS PGSSLDHPSSLDHPGS PGSSLDH	121 IdTRANKGQ FGRRILDEE FGRRILDEE WGNGATGLE VGQAGTADPD URTOGKAYL VAKGSIREN VAKGSIREN k Idd JLGQEQD5VGGI ILGQEQD5VGGI ILGQEQD5VGGI ILGQEQD5VGGI ILGQEQD5VGGI ILGQEQD5VGGI ILGGLP5DKPKLGG TLGGLP5NKPKLGG		131 139 GELQQLKIVCD GDIQQLSIIPG GDIQQLLTGD GDIQQLTTGD GSLQQLTVHPD FDLQQVHYCD GMISELKVRKT GFVESMKIILG GIMQDVQLVM GLUQDNHLIFD DSMDELKLVMG 1111111111 eeefffffff GELSEINLWNT GCLKDVILNGK GCLKDVILNGK GCLKDVILNGK GCLKDLFIDGQ GCLRDLFIDGQ
TspN PARPb COLSAIh COLIAAIh COLISAIh TSP3m NELg TSP4x IGNH ILIM CRP111 CRP111 CRP111 CRP111 APEXg LamG HSP62h ABP1r NEULAr GAS6m	81 RLALSVSGS-S' RVAVXKGQ-S' RIALSVHKK-T' RYAISVEKK-T' RFAMIVQGE-E' KLMLSVQG-S' TALLR.RGP5G RLSLAISAS-H HALSVAGE-S' ALLHLSGL5F: hhhhhh i: LCTTNSS2L' HICTTNSS2L' HICTSSS2AS HICTSSSSAS VVERSSAS VVERSSSAS VVERSSSAS TISVESLER-R TISVEELER-N	91 VTLYDCNK VTLILDCKK VTLILDCKK VTLILDCKK VTLILDCKK VTLVNCEB ASVHVDCSS ULIVYDCKL LILYVDCKL LILYVDCKL MILYVGCDI AKLYIDCNF MILYVGCDI AKLYIDCNF GSIQVDGE GSIQVDGE LLWVDGKE GSIQVDGE LLWVDGKE GSIQVDGE LLWVDGKE	101 	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVI PRSBRAIVDTRGIV PRSBRAIVDTGITV SSQLAPESSAGI PLGPRPMPPVGHVF ALAPIPAEVGGL PRPLSGIRLNTGSVW kkk ddd GLAVGRTLSQGELV LKKGYTVGAEASII SMATGHIVPEGGLL ULAPHPIKSGGVLI SASLADHPQLSMRIS SGSPNAVNAKGSIS SASLADHPQLSMRIS BLFQLERGLYHNLT	121 LGTRANKGQ FGRRILDEE FGRRILDEE FGTRILDEE KGNQATGLE LGLDAEQGK LGDAEQGKAL LGQRNAHG LGQRNAHG LGQRNAHG LGQRDAE VGG LGQRQDS VGG LGQEQDS VGG LGQEQDT VGG LGQEQDT VGG LGGAPDVATLTGG LGGLPFKKAGLVF VGGPFKSSELVQ		131 139 GELQQLKIVCD GDLQELSIFG GDLQQLLFVSD GDLQQLLFVSD GDLQQLTVHDD FDLQQVHIYCD GMXSELKVRKT GFVESMKIILG GFVESMKIILG GGLLQNIHLIFD DSMDELKLVMG GGLLQNIHHIFD GRLTGFNIWDS GGLLGVNIWMDF GRLTGFNIWDS GGLLSQFNIWDS GGLKDVILNGK GCVKNLVLHSA GCTRDLFTGQ GCMRDTFTGQ
TspN PARPb COL5A1h COL15A1h COL15A1h COL16A1h TSP3m TSP1c TSP4x 10NH 1LTM CRP11 CRP11 CRP1 1CM PTX3h APEXg LamG HSPG2h APE1 NEULAr GAS6m NEU3P	81 RLALSVSGS-ST RVAVAYKG2-ST RIALSVHKK-T RYAISVEKK-T RYAISVEKK-T RYAISVEKK-T RIALSHLARGFSG HILSALSAS-H NLTVQVTGS-N ALLLHLSGLSFT HLCTNNSELT TVVFTRSGR-M RVTALEGR-M RVTALEGR-M TISVEELER-M VVEPTRNGC-N	91 / //TLIYDCNK VTLIYDCKK VTLIYDCKK VTLIYDCKK VTLIVDCKK VTLIVDCKK VTLIVDCKK VTLIVDCKK VILYVDCKL SUVYTLI SUVYTLI SUVYTLI SUVYTLI VILYVNCE SUVYTLI SUVYTLI VILVVKKDA ATLQVDNK	101 	111 PRPFQGLIDVDGTIV PRSRAPVLDTRGVI PRSRATVDTNGITV PRSRATVDTNGITV PRSRATVDTNGITV PRSRATVDTNGITV PRSRATVDTNGITV PRPLSQLALPESSAGIT PRPLSQLALPESSAGIT NATVDTLSV LKKGYTVGAEASII PROSQLSLPRGLTP SPGPNVAVNAKGSII SASLADHQLSNRII PRSSLLDDDELN REFULERCLYNLNIZ REFULERCLYNLNIZ RCRQLTIPNTQAQII	121 FGRRIDEE FGTRIDEE FGTRIDEE VGQAGTADPD UGLDAEGK VGQAGTADPD	PFQ 	131 139 GELQQLKIVCD GDIQQELSIIPG GDIQQFLITGD GDIQQLTVHPD GDIQQLTVHPD GDIQQLVHPD GMISELKVRKT GFVESMKILG GINQDVQLLVM GLIQDNHLIFD DSMDELKLVMG 1111111111 GELSELNLWMT GGLSCHNWDS GGLSCHNWDS GGCLRDVINGK GCCRRDINLGH GGCLRDLFIDGQ GQCMSSWNLNG GQCNSCWNLNGK
TspN PARPD COL5Alh COL15Alh COL15Alh COL15Alh TSP3m NELg TSP1c TSP4x IGNH PTX3h APEXg LamC HSPG2h ABP1r NEULAT GAS6m NEU3DD LAMC3b	81 RLALSVSGS-S' RVAJSVEK-T' RYAJSVEK-T' RFANIVQGS-S' KLMLSVAGR-V' HALSVDGG-S' TALLR.RGP5G' NLTVQVTGE-N ALLLH.SGL5F; aaaaaaaaa2bi HUCTWSS22L' HUCTWTS22L' HUCTWTS22L' HUCTWTS22L' HUCTWTS22L' HUCTWTS22L' HUCTWTS22L' HUCTWTS2R-N RVTALREGR-R' PVELKNIGD-S' TISVELER-N VVRFRNGG-N	91 VTLYDCNK VTLILDCKK VTLILDCKK VTLILDCKK VTLILDCKK VTLIVNCEE ASVHVDCSS ULHVDCNK LILHVDCNK LILHVDCNK NLYVGCDI AKLYIDCNE ISLWVNGEI GSIQVDGEE GSIQVDGEE LLWVNCDA ATLQVDNWE SSLVVTGLK	101 	111 PRPFQGLIDVDGTIV PRSRAPVLDTRGVI PRSSRAIVDTRGIV PRSSRAIVDTGIV SSQLAPESSAGIF PLGPRPMPPVGHVF PLGPRPMDLPVGTFW PRPLSGIRLNTGSVW kkk dddd SIAVGRTLSQGGLL SIAVGRTLSQGGLL SMATGHIVPEGGLQ ULAPHPIKSGVLI SMATGHIVPEGGLQ SASLADHPQLSRGLYHINI SASLADHPQLSRGILD SASLADHPQLSRGILD SLPCNSTSIRAPV	121 LGTRANKGQ FGRRILDEE FGTRILDEE LGLDAEQGK VQAQGTADPD LIRTOQKAYL LIRTOQKAYL	PPQ	131 139 GELQQLKIVCD GDLQELSIFG GDLQQLLFVSD GDLQQLLFVSD GDLQQLTVHPD FDLQQUHVYCD GMLSELKVRKT GEVLQVHLVMG GLLQNIHLIFD DSMDELKLVMG GULGNVNMWDF GELSQFNIWDS GELSQFNIWDR GCCKRVLVLHSA GCCKRVLVLHSA GCCRRDWLGH GQLSGLYDGL GCLKNFQLDSK
TspN PARPb COL5Alh COL5Alh COL13Alh COL13Alh COL13Alh TSP3m NEL2 TSP4z 10NH 1LTM CRP11 CRP11 CRP11 CRP11 CRP11 CRP1 HSPG2h APE12 HSPG2h APE12 NEULAT MEUJAB NEUJAB NEUJAB	81 RLALSVSGS-ST RVAVAYKG-ST RIALSVHKK-T RYAISVEKK-T RYAISVEKK-T RYAISVEK-T RYAISVEG-ST TALLRLRGPSG TALLRLRGFSW LLLHLSGLSF, HLCTSWESA21 HLCTSWESA21 HLCTSWESA21 TVVFTSGR-M RVTALREGR-R RVTALREGR-R VTALREGR-R VVRTRNGC-S TVVFFROG-N TVVFFROG-N TVVFFROG-N	91 // //TLIYDCNK //TLIYDCKK	101 	111 PRPFQGLIDVDGTIV PRSARPVLDTRGVII PRSSHPMIDTNGIIV PRSSHPMIDTNGIIV PRSSRAIVDTNGITV PRSSRAIVDTNGIV PRSSRAIVDTNGIV PRPLSGLELERCAGLE PRPSGLELERCAGLE PRPSGLELERCAGLE PRSSGLE PRSSGLELERCAGLE PRSSGLE PRSSGLE PRSSGLE PRSSGLE PRSSGLE PRSSGLE PRSSGLE PRSSGLE PRSSGLE PRSSGLE PRSSGLE PRSSGLE PRSSGLE PRSSGLE PRS	121 LGTRANKGQ FGRRILDEE FGTRILDEE VGQAGTADPD VIRTGQKAYL VIRTGQKAYL VIRTGQKAYL VAKGSIEEN VAKGSIEEN K IdQEQDSVGG ILGQEQDSVGG ILGQEQDSVGG ILGQEDSNEVELJGG ILGQEDSNEVELJGG ILGGLDENKAGLVF ILGGLDENKAGLVF ILGGLDENKAGLVF ILGGLDENKAGLVF ILGGKDKGR LGSPPSKEVAJVA	PFQ 	131 139 GELQQLKIVCD GDIQQELSIIPG GDIQQFLITGD GDLQQLFVYDD GDLQQLTVHPD GFVESMKILG GINQDVQLLVM GLNQDVQLLVM GLNQDVLLVM GLNQDVLLVM GELSQENLWMT GGLSQENLWMT GGLSQFNIWDR GCLSQFNIWDR GCCKRDVLLSA GCCRRDWLGH GCCRRDWLGH GCCLKNFQLDSK GSLRLDVNNQ
TspN PARPD COL5Alh COL15Alh COL15Alh TSP3m TSP1c TSP4x TSP4x IGNH ILIM CRP111 CRPh PTX3h APEXg HSP62h ABPC2h MEULAr GAS6m NEULAR AGRNc SLIT2d	81 RLALSVSGS-ST RVAJSVEK-T RYAISVEK-T RPANIVGG-ST TALLRRGP5G HILLRGP5G RLSLAISAS-H NLTVQVTGE-N ALLLHSGL5F: Aaaaaaaa2b HUCHTWSSV2E HICTSWESA2T HLCTSWESA2T HLCTSWESA2T HLCTSWESA2T HLCTSWESA2T ST SVELKANGD-S TISVELER-N TVVFTRGG-N TVVFGHDGE-K QLVVNRNR-S SU	91 // // TLIYDCNK // TLIDCKK // TLIDCKK // TLIDCKK // TLIDCKK // TLIYDCKK // TLIYDCKK // TLIYDCKK // TLIYDCKK // TLIYDCK // TLIYDCKK // TLIYDCK // T	101 	111 PRPFQGLIDVDGTIV PRSRAPVLDTRGVI PRSEATVDTNGIV PRSEATVDTNGIV SSQLAPESSAGIF PALAPIPAEVSGLE PREVPMDLVGTFW BEFPHQLKAENSRMY PRPLSGIRLNTGSVH LKKGYTVGAEASII SMATGHIVPEGGILC VLAKGYTVGAEASII SMATGHIVPEGGILC VLAKGYTVGAEASII SASLADHPLKSGGVLI PGSSLLDLDDELS LFQLERCLYHLNT SLFQLERCLYHLNT SLFQGSTISIRAPVI SPTTDGLNLDTDLI SPTTDGLNLDTDLI	121 IdTRANKGQ FGRRILDEE FGRRILDEE VGQAGTADPD UGLDAEGK VGQAGTADPD URT0GKAYL VAKGSIREN URTLQK K IdGQEQDSFGGI IGQEXNGCC-VGGI IGQEXNGCC-VGGI IGGEXNGCCX IGGEXNGCCX IGGEXNGCCX IGGEXNGCCX IGGEXNGCX IGGX IGGX IGGX IGGX IGGX IGGX IGGX IGGX IGGX IG		131 139 GELQQLKIVCD GDIQQELSIIPG GDIQQLFVSD GDIQQLFVSD GSLQQLTVHPD FDLQQVHYCD GMSELKVRKT GEVESMKIILG GFVESMKIILG GELQNIHLIFD DSMDELKLVMG 1111111111 eeefffffff GELSQFNWDR GCLKDVILNGK GCCKEVNLVLHSA GCCKEVNGLG GCKEVWINHK
TspN PARPb COLSAIH COLIAIH COLISAIH COLISAIH COLISAIH TSP3m TSPic TSP4x ISPI CRP111 CRP111 CRP111 CRP111 CRP111 CRP114 HSPC21 HS	81 RLALSVSGS-ST RVAVAYKG-ST RIALSVHKK-N RVAISVEKK-T RFANIVGE-E RIALSVLKG-ST TALLRLRGPSG TALLRLRGPSG NLSLAISAS-H NLTVQVTGS-N ALLLHSSQLE HICTSVESSQLE HICTSVESSQLE HICTSVESSGAS VVETKSGR-R PVELKNNGD-ST HVDFQRDGS-ST TVVFGHDGE-K TVVFGHDGE-K VVEFTNGG-N TVVFGHDGE-K VVETKNGA-ST AVELLAIKK-N	91 // ////////////////////////////////	101 	111 PRPFQGLIDVDGTIV PRSRAPVLDTRGVII PRSSHPMIDTNGITV PRSSHPMIDTNGITV PRSSRAIVDTNGITV PRSSRAIVDTNGITV PSSQLAPESSAGIT PALAPIPPAEVSGLE PSSQLAPESSAGIV PRPLSGIRLNTGSVF kkk dddd IAVQRTLSQGLVV LKKQTYQAEASII SMATGHIVPSGGLQ LAPWHPIKSGGLV LAPWHPIKSGGLV LSKQTYQAEASII SASLADHPQLSRGIF SASLADHPQLSRGIF SLPGDERGLVHLNII SLPGENIVLNINGVI SLPGDERGLVHLNII SLP	121 LGTRANKGQ FGRRILDEE FGTRILDEE FGTRILDEE VGQAGTADPD UGQDXGTADPD LIGQDXAYL LIGQRNAHG VAKGSIREN K IGQDXGC VGQD VGGD FKSELVQ LIGQEXNGCC VGGI FKESELVQ LIGGLEPSKALRF LIGGELLPTSKLRF //GGLPENKASLV VGGJ FKESELVQ LIGGKDKGR LIGSPSGCHSLP VGGAPEDQMAVVA /LGGLPVDPAQAYVA /LGGLPVDPAQAYVA /LGGLPVDPAQAPILE	PFQ 	131 139 GELQQLKIVCD GDIQQELSIIPG GDIQQFLTGD GDIQQFLTGD GDLQQLFVYD GBLQQLTVHPD FPLQQVHYCD GGVSSKILG GIVQDVQLLVM GLVQNHLIFD DSNDELKVVMG 11111111111 eeefffffff GGLSELNUMNT GGLSQFNIWDR GGLSQFNIWDR GCUKNQLHSA GCUKNULHSA GCUKNQLDFIDGQ GCMRSWNNLMG GQCLSVFUGSL GGCLNFQGLDSK GGLSVFUGGR
TspN PARPb COL5A1h COL15A1h COL15A1h COL16A1h TSP3m TSP1c TSP4x 1GNH 1LTM CRP111 CRP14 ILTM CRP111 CRP14 JCR17 MEULAT GAS6m NEULAT ASPEN ASPEN LAMA3h AGRNC SLIT2d FATd	81 RLALSVSGS-ST RVAVAYGQ-ST RIALSVHKK-T RFAMIVQGE-ST TALLRRGP5G KLSLATSAS-H MLTQVTGE-N ALLLHLSGLSF HLCTNSSZL HUCTTNSSZL HUCTTNSSZ TVVFTRSGR-N RVTALREGR-N RVTALREGR-N TVVFRSG-N TVVFRGG-N TVVFRHGG-ST TVVFRHGS-ST VVRPTRNG-N VVRTRNG-ST VVRPTRNG-N VVRTRNG-ST VV	91 / //TLIYDCNK VTLIYDCNK VTLILDCKK VTLILDCKK VTLIVDCK VTLIVDCK VTLIVDCK LOLYVDCH LLHVDCNK XILYVDCH AKLYIDCNF IIIII DbDbbbb cc ALIYDCNF TLAVDGH GTLSVDGLE GTLSVDGLE GTLSVDGLE GTLSVDGLE GTLSVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE GMLAVDGLE MACADSGL GMLAVDGLE MACADSGL GMLAVDGLE MACADSGL GMLAVDGLE MACADSGL GMLAVDGLE MACADSGL MACADSGL GMLAVDGLE MACADSGL MACADSG	101 	111 PRPFQGLIDVDGTIV PRSRAPVLDTRGVI PRSSRATVDTNGIV PRSSRATVDTNGIV SSQALAPESSAGIF PALAPIPAEVSGLE UGPRRPMRPVGTVF PRSQGLELBRGAGLE SSQGLELBRGAGLE SSQGLELBRGAGLE SSQGLELBRGAGLE SSQGLELBRGAGLE SSQGLELBRGAGLE NATGHIVPSGGIL LRGYTVGABASII SMATGHIVPSGGILG LSQGLVINAKGSII LPGNSQLSLPRGL9 SGENVAVNAKGSI SLPQLERGLYHLNI SLPQERSTISIRAPV SPTGTDGLNDTDLI BGSNDYLKLTTPMI VGPVGTLMFNRQPLI	121 LGTRANKGQ FGRRILDEE FGTRILDEE FGTRILDEE VGQAGTADPD ILGLDAEGGK VGQAGTADPD ILGQENGC K LGQEQGS FGG ILGQEQDS FGG ILGQEQDS FGG ILGQEQDS FGG ILGQEQS FGG ILGQEQS FGG ILGQEQS FGG ILGQEQC VGG ILGQEQC VGG ILGQENKGC CGGLPSDKPKLGG ILGGLPENKAGLVP VGGIPFKESELVQ ILGGLPENKAGLVP VGGAPEDCMAVVA VLGGLPEDCMAVVA ILGGLSSADPILER ILGGLSSADPILER		131 139 GELQQLKIVCD GDIQQELSIIPG GDIQQFLITGD GDIQQLFUYDD GDIQQLTVHPD FDLQQVHYCD GMSDELKUVKT GELQUNHLIFD DSMDELKLVMG GELSQLNUMDF GGLSQFNUMDF GGLSQFNUMDF GGCTRRDIKDMDF GGCTRRDIKDMDF GGCTRRDIKDK GCCKRUVLKSA GGCKEVWINKK GGCKEVWINKK GGCKEVWINKK
TspN PARPD COLISAIH COLISAIH COLISAIH COLISAIH TSP3m NELG TSP1c TSP4x IGH PTX3h APEXg LamG HSPG2h ABP1r NEULAT GAS6m NEU3BD LAMA3H AGRNC SLIT2d FATA	81 RLALSVSGS-S' RVAJSVEK-T' RYAISVEK-T' RFANIVQGS-V KLMLSVAGR-V' HALSVDGG-S' TALLR.RGP5G RLSLAIAS-H MLTVQVTGE-N ALLLH.SGL5F; aaaaaaaaa2b HUCTMTS22L HLCTTMTS22L HLCTTMTS22L HLCTTMTS22L HLCTTMTS22L HLCTTMTS22L HLCTTMTS22L HLCTTMTS22L HLCTTMTS22L HLCTMTSS22L HLCTMTSS22L HLCTMTSS2L RVTALREGR-R PVELKNMGD-S' TVSTBLER-N VVRFRNGG-N VVRFRNGG-S' TVVTRNGR4S TVKTDVVRR.K-S KVTATNRG4S	91 VTLYDCNK VTLIDCKK VTLIDCKK VTLIDCKK VTLIDCK VTLIDCK VTLIDCK VTLVNCEE ASVHVDCSS ASVHVDCSS LLHVDCNK LLHVDCNK MEXPUDGK GSIQVDGEE GSIQVDGEE GSIQVDGEE GSIQVDGEE GSIQVDGEE GNLVVDGL GMLAVDGEF FTLRVDREI VMKCADSG GFITVGRE	101 	111 PRPFQGLIDVDGTIV PRSPRPVLDTRGVI PRSSRATVDTRGIV PRSSRATVDTGITV SSQLAPESSAGIF PLGPRPMPVGHVF ALAPI PPABVSGL PFYEQLARENSRY PRPLSGIRLNTGSV kkk dddd SIAVGRTLSQGLVV LKKQYTVGAEASI SMATGHIVPEGGILQ LLKQYTVGAEASI SMATGHIVPEGGILQ LLAPHPIKSGVLI SPGPVLAVNAKGSTI SPGPVLAVNAKGSTI SPGPVLAVNAKGSTI SLPONSTISIRAPV SPGSDLLDDDLD SPTTJGLNLDTDLI NEGSNDLLKTTPMI VGPVGTLMFNKQPLB VVGPQGTLMFNKQPLB	121 LGTRANKGQ FGRRILDEE FGTRILDEE LGLDAEQGK UQQAGTADPD LGLDAEQGKA VAKGSIREN VAKGSIREN VAKGSIREN	PPQ	131 139 GELQQLKIVCD GDLQELSIIPG GDLQQLLYVSD GDLQQLLYVSD GDLQQLTVHPD FDLQQUHYCD GMLSELKVRKT GEVLQVIHYCD GMLSELKVRKT GULQNIHLIFD DSMDELKUVMG GULQNIHLIFD DSMDELKUVMG GGLCNIWMDF GGLSUVINWDF GGLSUVINWDF GGLSUVINGK GCVKDVLINGK GCVKDVLINGK GCVKDVLNGK GGCKNFQLDSK GGCKFVDGL GGCKFVDL GGCKFVDGL GGCKFVDL GGCKFV
TspN PARPb COL5Alh COL5Alh COL13Alh COL13Alh TSP3m TSP1c TSP4x 1CNH 1LM CRP11 CRP11 CRP11 CRP11 CRP11 CRP1 LamG HSPG2h ASP1r NEULAr GAS6m NEULAR GAS6m NEULAR SLIT2d FATG LAMA1h	81 RLALSVSGS-ST RVAVAYGQ-ST RIALSVHKK-T RYANIYQGE-ST ALLRURGP50 HILSUNGG-ST ALLLRURGP50 ALLLHLSGL5F, ALLLHLSGL5F, HICTSWESA2T HICTSWESA2T HICTSWESA2T HICTSWESA2T TVVFTSGR-M RVTALRGR-R TVVFTSGR-M TVVFRSGR-M TVVFRSGR-M TVVFRSGR-M TVFTSGR-ST ISSUELER-M VVRFTRNGG-S TVFTDVKR-K TVTTDVKR-K TVTTDVKR-K TVTTDVKR-K TVTTDVKR-K	91 / //TLIYDCNK //TLIYDCKK //TLIYDCKK //TLIYDCKK //TLIYDCKK //TLIYDCKK //TLIYDCKK //TLYYDCEE //LIYYDCK //LIYYDCK //LIYYDCK //LIYYDCK //LIYYO	101 	111 PRPFQGLIDVDGTIV PRSRAPVLDTRGVI PRSSRATVDTNGITV PRSSRATVDTNGITV PRSSRATVDTNGITV SSQALAFESSAGIT PRSSRATVDTNGITV SSQGLELERGAGLE PALAPIPAEVSGLE PRFVSQLKAENSRW PRPLSGIRLNTGSVE KKGYTVGAEASII MATGHIVPEGGILC LFGNSQLSLPRGLFS SFGPNVAVNAKGSII SSGLAUPHOLSMRII SGGRUTIFNTQAQIJ SLPQUERGLVHIMII KGRQLTIFNTQAQIJ SLPQUERGLVHIMII KGRQLTIFNTQAQIJ SLPQUERGLVHIMII KGRQLTIFNTQAQIJ SLPQUERGLVHIMII FGSDVLKLTFPMI GPVGTLNFNKQPLD PRSSBUTLKLTPPMI FFKASHLKVPYGLFT PRSSBLVKFJFT	121 LGTRANKGQ FGRR ILDEE FGTR ILDEE FGTR ILDEE FGTR ILDEE FGTR ILDEE FGTR ILDEE FGTR ILDEE GGAPDAL ILGLDAEGGAS K ILGQEQDS FGGI ILGQEQDS FGGI ILGQEQDS FGGI ILGQEQDS FGGI ILGQEQDS FGGI ILGQEQDS VGG FLGGLPSDKPKLGG FLGGLPSDKPKLGG ILGGLPSDKPKLGG ILGGLPSDKPKLGG ILGGLPSDKPKSLVQ ILGGLPSDKFKSLVQ ILGGLPDPSGKPKSLP ILGGLPSQVQAKTI IGGLSSADPILEE ILGGLPSQVQAKTI	PFQ 	131 139 GELQQLKIVCD GDIQQLSIIPG GDIQQFLITGD GDLQQLFVYD GDLQQLTVHPD FDLQVHYCD GMSDEQLVW GHVSWKILG GINQDVQLVW GLUQNIHLIFD DSMDELKVWG 1111111111 eeefffffff GGLSSLHUMT GGLSQLWMDS GGCLRNUKLSA GCCKNIKVLNG GCCKNIKVLNG GCCKNIKVLNG GCCKNIKVNN GCCLSVIGGR %CCGNUNNSK GCCLSALLNGR

Figure 5. Multiple alignment of the common conserved region of the proposed TspN/LamG/pentraxin Membership superfamily. 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) 1LIM (theoretical model, and 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 polyphe*mus*; m, mouse; r, rat; x, Xenopus laevis): PARP, proline and argininerich protein; COLxAy, alpha y chain of collagen x; TSP, thrombospondins; CRP11, C-reactive protein 1.1; PTX3, pentraxin-like protein PTX3; APEX, acrosomal pentraxinlike 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 Ca²⁺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 Ca²⁺-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 PRO-SITE 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). Steroidbinding and dimerization domains of human sex hormone-binding globulin partially overlap: steroids and Ca²⁺ 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 EGFlike 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 noncollagenous 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 egflike 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)