Domains in plexins: links to integrins and transcription factors

Integrins are adhesion molecules that bind diverse cell-surface and extracellular-matrix ligands[1]. They are heterodimeric receptors, containing α and β subunits. No significant sequence similarity between the extracellular domains of integrin β subunits and any other protein has been reported, although the presence of a DXSXS motif, and secondary-structure predictions, suggests that the most-conserved region adopts an I-domain-like fold (also called a von Willebrand factor A domain)[2–4]. Furthermore, the C-terminal third of the extracellular domain contains four internal, cysteine-rich repeats[5]. The only other cysteine-rich region (a cluster of seven cysteine residues in 50 residues) is in the N-terminal segment of mature integrin β subunits. Six of these cysteines form disulfide bonds to one another; the remaining, first, cysteine forms a long-range disulfide bond that links the N-terminal and C-terminal cysteine-rich regions[6].

PSI-BLAST searches[7] with the N-terminal cysteine-rich region of individual integrin β subunits retrieve a region in a mouse neuronal cell-surface molecule, plexin 2 (Ref. 8), at the third iteration ($E = 3 \times 10^{-4}$). Further iterations reveal homology to previously described internal repeats of this region in plexins[8,9]. Subsequent PSI-BLAST searches with this repeat reveal in the first iteration significant homology to a family of proteins that act as 'semaphores' for growth-cone guidance[10], the semaphorins ($E = 3 \times 10^{-4}$). In later iterations, proteins related to a signaling receptor, mahogany, that functions in the brain as a suppressor of obesity are retrieved[11] (E values are of the order of 10^{-4}). Previously, it has been shown that the three repeats in plexin are homologous to a small region of the hepatocyte growth factor (HGF) receptor, MET (Ref. 9), and also to the virus-encoded semaphorin receptor (VESPR)[12]. Reciprocal studies using both sequence profiles[13] of the repeats from the plexin family and PSI-BLAST searches with several members of the major subfamilies indicate that other proteins (such as MEGF8 or C21orf1) contain the repeat (Fig. 1). We named this region the PSI domain (after the better-characterized families plexins, semaphorins and integrins). The PSI domain is part of the original definition of the sema domain[10], but further studies of the modular architecture of the semaphorins reveal that there are indeed semaphorins that lack the PSI domain, such as A39R from vaccinia virus. This leads to a redefinition of the original sema domain, and only this redefined sema domain seems to be a marker for semaphorins.

The PSI domains of the proteins shown in Fig. 1 are ~50 residues in length and usually contain eight cysteine residues. On the basis of experimental determination of disulfide bonds in integrins[6], we can predict the location of disulfide bonds (between cysteine two and cysteine four, between cysteine three and cysteine eight, and between cysteine five and cysteine seven) for the entire family (Fig. 2a). The remaining two cysteines (cysteine one and cysteine six), when both are present, are also predicted to be disulfide bonded; in integrins, cysteine one forms the long-range disulfide bond[6], and cysteine six is absent. In several family members, cysteine five and

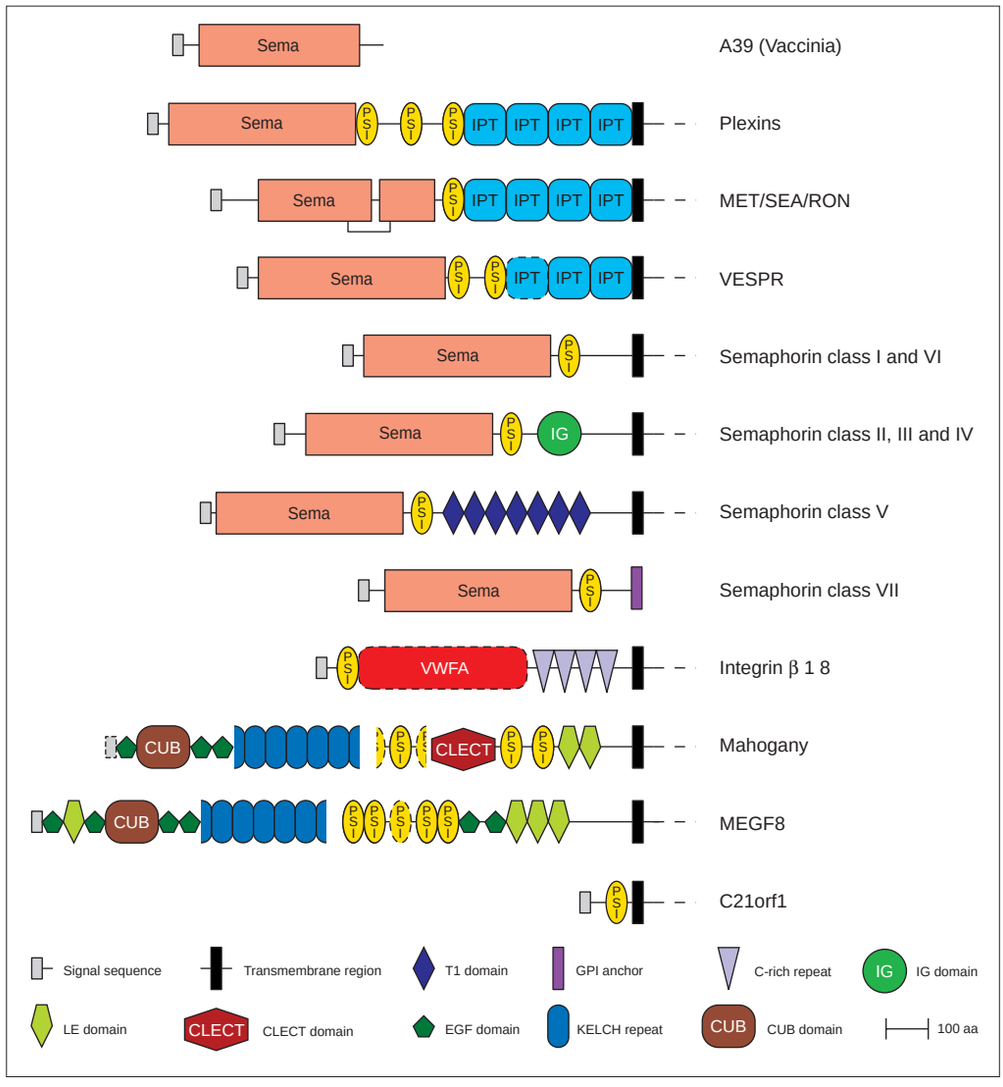

Figure 1

The extracellular region of proteins containing PSI and IPT domains. Only regions of proteins that have distinct modular organizations to their extracellular regions are shown. Domain names are according to Ref. 21. A broken line surrounding a domain indicates that it does not give significant *E* values, but its presence is supported by context information. The cysteine (C)-rich repeats in the integrins somewhat resemble epidermal growth factor (EGF)-like domains, although they contain two additional cysteines. aa, amino acid residues; GPI, glycosylphosphatidylinositol; IG, immunoglobulin; LE, laminin epidermal growth factor-like; VESPR, virus-encoded semaphorin receptor; VWFA, von Willebrand factor type A domain.

0968–0004/99/$ – See front matter PII: S0968-0004(99)01416-4

(a)

```
                           1       2 3           4 5      6                 7                   8
semaF        hs  497  RCQFYR---TRST-CIGAQDPY----CGWDVV---MKKCTSL---------EESLSMTQWEQSIS-------ACP  U52840
semaIII      hs  510  RCDIYG---KACAECCLARDPY----CAWD-----GSACSRY---------FPTAKRRTRRQDIRNG-DPLTHCS  D49423
semaI        tc  484  HCASK----TRCKDCVELQDPH----CAWDAK---QNLCVSI---------DTVTSYRFLIQDVVRG--DDNKCW  A49423
semaK1       hs  460  LCEVYG---GGCHGCLMSRDPY----CGWD-----QGRCISI-----------YSSERSVLQSINPA-EPHKECP  AF071542
semaG        mm  502  FCEKH----GSCEDCVLARDPY----CAWSPA---IKACVTL--------HQEEASSRGWIQDMSG---DTSSCL  U69535
PLXN1_a      mm  518  SCVQY----TSCELCLGSRDPH----CGWCVL---HSICSR---------QDACERAEEP-QRFAS---DLLQCV  P70206
VESPR_a      hs  460  NCNKH----KSCSECLTATDPH----CGWCHS---LQRCTF---------QGDCVHSENL-ENWLDISSGAKKCP  O60486
K04B12.1_a   ce  486  TCSHH----SSCTECLVSVDPL----CQWCHP---TQSCTT---------SARCTS------PVTS------QCP  O45657
PLXN1_b      mm  658  NCSVH----QSCLACVNGSFP-----CHWCKY---RHVCTN--------NAADCAFLEGR-VNMSE------DCP  P70206
VESPR_b      hs  591  NCSSL----KECPACVETG-------CAWCKS---ARRCIH--------PFTACDP-----SDYER---NQEQCP  O60486
K04B12.1_b   ce  617  DCSGY----GTCSSCMSSEYN-----CAWCSG---LHKCSN-----------SC-------GALEKSK----ACV  O45657
PLXN1_c      mm  806  KCPALR---QSCGLCLKADPRF[ 1]CGWCVA---ERRCSL---------RHHCPADSP--ASWMHAHHGSSRCT  P70206
K04B12.1_c   ce  747  SCTNLA---SDCSSCLALSPSL[ 1]CGWC-----NRQCSH-----------ECHE-----SKATA------VCD  O45657
MET_HUMAN    hs  519  GCRHF----QSCSQCLSAPPFV[ 1]CGWC-----HDKCVR---------SEECLS-----GTWTQ-----QICL  P08581
RON_HUMAN    hs  526  GCRHF----LTCGRCLRAWHFM[ 1]CGWC-----GNMCGQ---------QKECP------GSWQQ-----DHCP  Q04912
MEGF8_a      hs  768  PCRLL----SSPEACNQSGA------CTWCHG-----ACLS---[ 3]AHRLGCG---------------GSPCS  D1033432
MEGF8_b      hs  816  ECRRL----RTCSECLARHPRT[13]CKWCT------NCPEG----------ACIGR-------------NGSCT  D1033432
MEGF8_c      hs  962  PCHLL----PNCTSCLDSKGAD[ 5]CVWSSS---LQQCLSP--------------------SYLPL-----RCM  D1033432
MEGF8_d      hs 1012  GCAQA----TQCALCLRRPH------CGWCAW[5]GGRCMEGGL[ 3]RDGLTCGRPG---ASWAFL-----SCP  D1033432
mahogany_a   hs  702  RCDQH----TDCYSCTANTND-----CHWC-----NDHCVP--------RNHSCSEGQISIFRYE-------NCP  AF116897
mahogany_b   hs  931  PCALR----TACGDCTSGSSE-----CMWCSN---MKQCVDSNA[ 4]FPFGQC-------MEWYT----MSTCP  AF116897
mahogany_c   hs  985  NCSGY----CTCSHCLEQPG------CGWCTD[5]KGKCIEGSY[20]LNSSMCLE--DSRYNWSFI-----HCP  AF116897
C37C3.7      ce    9  YCRIGSNDLNTCETCVGKGSN-----CFWCGG[1]TKRCMPF--[ 1]WYYPDCNI---KHVKYN-------VCW  Q22913
C21ORF1      hs   39  ACSQNTN--KTCEECLKNVS------CLWCNT---NKACLDYPV[ 4]PPASLCKL---SSARWG-------VCW  P53801
ITB1_HUMAN   hs   26  RCLKANA--KSCGECIQAGPN-----CGWCT----NSTFLQEGM----PTSARCD----DLE-ALK----KKGCP  P05107
ITB2_HUMAN   hs   24  ECTKFKV--SSCRECIESGPG-----CTWCQ----KLNFTGPG-[ 1]PDSIRCD----TRPQLLM-----RGCA  P18564
ITB3_HUMAN   hs   30  ICTTRGV--SSCQQCLAVSPM-----CAWCS----DEALPLGS--------PRCD----LKENLLK-----DNCA  P26012
ITB4_HUMAN   hs   29  RCKKAPV--KSCTECVRVDKD-----CAYCT----DEMFRD----------RRCN----TQAELLA-----AGCQ  P05106
ITB6_HUMAN   hs   23  LCLGGA---ETCEDCLLIGPQ-----CAWCA----QENFTHPS-----GVGERCD----TPANLLA-----KGCQ  P05556
ITB8_HUMAN   hs   46  RCASSNA--ASCARCLALGPE-----CGWCV----QEDFISGG-----SRSERCD----IVSNLIS-----KGCS  P16144
consensus(80%)        .Ct.h....ttCttChtttt......CtWC......ttC.t...........tC........th........tC.
sec. struct. (pred.)  ..........................................................hhhhhhh.........
```

(b)

```
SEX_HUMAN_a   hs   839   HPRITQIHPLVGPKEGGT-RVTIVGENLGL-----LSREVGLRVAG------VRCNSIPAEYI---------SAERIVCEMEESLVPSP--
SEX_HUMAN_b   hs   934   TPTFDQVSPSRGPASGGT-RLTISGSSLDA-----GSRVTVTVRD-------SECQFVRRD------------AKAIVCISPLSTL-----
SEX_HUMAN_c   hs  1022   DPTVTRLEPTWSIINGST-AITVSGTHLLT-----VQEPRVRAKYRGIET-TNTCQVIN--------------DTAMLCKAPGIFLGRPQ-
SEX_HUMAN_d   hs  1124   DPSFEPLGPSGVLDVKPGSHVVLKGKNLIP-----AAAGSSRLNYTVLIG-GQPCSLTV-------------SDTQLLCDSPSQTGRQPV-
MET_MOUSE_a   mm   561   LPAVYKVFPTSAPLEGGT-VLTICGWDFGFRKNNKFDLRKTKVLLG-----NESCTLTLSES----------TTNTLKCTVG--------
MET_MOUSE_b   mm   655   DPVITSISPRYGPQAGGT-LLTLTGKYLNS-----GNSRHISIG-------GKTCTLKSV------------SDSILECYTPAQT------
MET_MOUSE_c   mm   740   DPVVYEIHPTKSFISGGSTITGIGKTLNSV----SLPKLVIDVHEVGVNY-TVACQHR--------------SNSEIICCTTPSLKQ----
MET_MOUSE_d   mm   837   NPVFEPFEKPVMISMGNENVVEIKGNNIDP-----EAVKGEVLKVG-----NQSCESLHWH------------SGAVLCTVPSDLLK----
RON_HUMAN_a   hs   568   PPKLTEFHPHSGPLRGST-RLTLCGSNFYLHPSGLVPEGTHQVTVG-----QSPCRPLPKDSSKLRPVPRKDFVEEFECELEPLGTQ----
RON_HUMAN_b   hs   683   EPVLIAVQPLFGPRAGGT-CLTLEGQSLSV-----GTSRAVLVNG-------TECLLARV------------SEGQLLCATPPGA------
RON_HUMAN_c   hs   769   DPVVLSISPNCGYINS---HITICGQHLTS-----AWHLVLSFHDGLRAV-ESRCERQL--------------PEQQLCRLPEYVVRDPQG
RON_HUMAN_d   hs   874   KPEEHAIKFEYIGLG----AVADCVGINVT-----VGGESCQHEFRGD---MVVCPLPPSLQLG------QDGAPLQVCVDGECH------
VESPR_a       hs   753   DPRFTGYRVESEVDTELEVKIQKENDNFNISK--KDIEITLFHGENG----QLNCSFENITRN--------QDLTTILCKIKGIKTASTI-
VESPR_b       hs   843   LPHCSLIFPATTWISGGQ-NITMMGRNFDV-----IDNLIISHELKGNINVSEYCVA------------------TYCGFLAPSLKS---
YMS5_CAEEL    ce   460   SNGGMMTAPPWGPMHGGM-AINVSGPCLRP-----ADSVKVNFENW-----QTSCTRLSR--------------VRARCIMPMFH------
OLF-1         ce   268   VPVIKALFPSEGWIQGGTQVVLIGENFFEG-----LQVAFGTASPNWG----ESVQLISP--------------HAIRVTTPPKHSA----
KBF2_CHICK    gg   221   NLKISRMDKTAGSVRGGDEVYLLCDKVQKD----DIEVRFYEDDENG---WQASGDFSPTDVH---------KQYAIVFRTPPYHK-----
NFATx2        hs   599   LPHIEKYSINSCSVNGG-HEMVVTGSNFLP----ESKIIFLEKGQDGRPQWEVEGKIIREKCQG----------AHIVLEVPPYHN-----
XCOE2         xl   250   TPCIKAISPSEGWTTGGAMVIIIGDNFFDG-----LQVEFGTMLVWS-----ELMTP-----------------HAIRVQTPPRHIP----
1CYG          bs   494   TPIIGHVGPMMGQVG---HQVTIDGEGFG------TNTGTVKFG-------TTAANVVSWSN------------NQIVVAVPNVS------
consensus (80%)          ..hh.tht.....ht....hhhht.t.h.......ht.hhhtht.........Cth................tthhCth...h......
sec. struct.(1CYG)       .EEEEEEE.EEEEEE...EEEEEEEEE...........EEEEEE.......EEEEEEEEEEE............EEEEEEEE.........
sec. struct.(pred.)      ..eEEEE...........EEEEE..............EEEEEE..........eEEEE.................eEEEE...........

SEX_HUMAN_a           -PPGPVEL----CVGD-CSADFRTQ-----SEQVYSFV        P51805
SEX_HUMAN_b           -GPSQAPI----TLAI-DRANISSP-------LIYTYT        P51805
SEX_HUMAN_c           -PRAQGEHPD--EFGF-LLDHVQTA--RSLNRSSFTYY        P51805
SEX_HUMAN_d           -MVLVGGL----EFWL-GTLHISAE-------RALTLP        P51805
MET_MOUSE_a           -PAMSEHF----NVSV-IISNSRE-------TTQYSAF        P16056
MET_MOUSE_b           -TSDEFPV----KLKI-DLANRET--------SSFSYR        P16056
MET_MOUSE_c           -LGLQLPLK--TKAFF--LLDGIL-------SKHFDLT        P16056
MET_MOUSE_d           -LNSELNIEWKQAVSSTVLGKVIVQ-------PDQNFA        P16056
RON_HUMAN_a           -AVGPTNVSLT-VTNMPPGKHFRVD-----GTSVLRGF        Q04912
RON_HUMAN_b           -TVASVPL----SLQV-GGAQVPGS-------WTFQYR        Q04912
RON_HUMAN_c           WVAGNLSARGDGAAGF-TLPGFRFLPPPHPPSANLVPL        Q04912
RON_HUMAN_d           -ILGRVVR----PGPD----GVPQ----------STLL        Q04912
VESPR_a               -ANSSKKV----RVKL-GNLELYVE--QESVPSTWYFL        O60486
VESPR_b               -SKVRTNV----TVKL-RVQDTYLD-------GTLQYR        O60486
YMS5_CAEEL            -KIGLVPIRMS-RDGG-QSFPFFGK--------FYVVN        P34501
OLF-1                 -GPVDVTLQY--KSKT-YSRGTPL---------RFSYI        O61576
KBF2_CHICK            -PKIDRPVTVFLQLKRKRGGDVSDS-------KQFTYY        P98150
NFATx2                -PAVTAAVQV--HFYL-CNGKRKK-----SQSQRFTYT        U85428
XCOE2                 -GVVEVTLSY--KSKQ-FCKGAPG---------RFIYT        O57515
1CYG                  PGKYNITVQS--SSGQ-TSAAYD----------NFEVL        1CYG
consensus (80%)       .....hth.h...ht.....th............hthh
sec. struct.(1CYG)    EEEEEEEEEE..EEEE.EEEEEE..........EEEEE
sec. struct.(pred.)   ....eeEEEE..EEe....................eeee.
```

Figure 2

Multiple alignments of selected members of the PSI- and IPT-domain families. The names of the proteins (multiple domains in the same proteins are labeled a, b, c or d), the species and the number of the residue at the start of the domain are shown on the left. Database accession numbers are shown on the right. Conserved cysteine residues are shown in red; the conserved tryptophan residue is shown in blue; conserved hydrophobic residues are shown in green; other conserved residues are shown in bold. The consensus sequence shown below contains conserved features in the domain: C and W denote conserved cysteine and tryptophan residues; t and h indicate turn-like/polar and hydrophobic positions, respectively. Predicted secondary structure (sec. struct. pred.)[14] is also shown (h, helix; E, β-sheet predicted with high significance; e, β-sheet predicted with lower significance). **(a)** Multiple alignment of different PSI domains. Cysteine residues are numbered above the alignment and color-coded on the basis of predicted disulfide-bonded residues[6]. The predicted secondary structure of the PSI domain is a consensus derived from independent results for plexins, semaphorins, integrins and mahogany-like domains. **(b)** Multiple alignment of different IPT domains. The known secondary structure of the D domain of *Bacillus stearothermophilus* (bs) cyclodextrin glucanotransferase (1CYG) is shown below the alignment. ce, *Caenorhabditis elegans*; gg, *Gallus gallus*; hs, *Homo sapiens*; mm, *Mus musculus*; tc, *Tribolium confusum*; xl, *Xenopus laevis*.

cysteine seven are both missing, which provides additional support for the idea that the two residues are normally bonded together (Fig. 2a). Secondary-structure predictions[14] strongly indicate that an α-helix is present between cysteine seven and cysteine eight (Fig. 2a) and thus support the alignment shown in Fig. 2a.

The presence of multiple PSI domains, and of a (redefined) sema domain, classifies plexin, MET and VESPR (Fig.1) as semaphorins. Moreover, it leaves only the region N-terminal to the transmembrane region in the extracellular part of plexins undescribed. PSI-BLAST searches with this region suggest that it has a repeat-like character. For example, if a search is initiated with the second repeat of RON_HUMAN (Fig. 2b), not only three internal repeats but also regions in MET and SEX are significantly similar (E values are 10^{-10} and 10^{-6}, respectively). Further iterations reveal homology to Olf1/Ebf-like transcription factors (e.g. for OLF-1, $E = 10^{-4}$) and to the nuclear factor of activated T cells (NFAT) family of transcription factors. Wisetool profiles[13] and Hidden Markov Model[15] searches based on the three repeats of the motif in plexins and MET detect a fourth repeat in these proteins, two repeats in VESPR and one in the hypothetical protein YMS5_CAEEL (Fig. 2b). A BLAST search with the motif of the transcription factor XCOE2, a close homolog of OLF-1, retrieves the immunoglobulin-like (D) domain of *Bacillus stearothermophilus* cyclodextrin glucanotransferase, whose structure is known (PDB accession number: 1CYG), with a significant E value ($E = 3 \times 10^{-4}$). We therefore propose the name IPT (for immunoglobin-like fold shared by plexins and transcription factors) for this domain (Fig. 2b).

Within the semaphorin family, the IPT domain is present only in plexins, MET and VESPR, but the PSI domain is associated with many more family members. Although most seem to have a role in neuronal development, several family members appear to have immunological functions[16]. Mahogany[11] is the only other PSI-domain-containing protein for which functional information is currently available. It is a signaling receptor and functions in the brain as a suppressor of obesity[11]. A receptor-like architecture for most of the proteins that contain the PSI domain is thus the only clear common theme, although soluble forms exist – at least in the cases of some semaphorins[16] and mahagony[11]. A monoclonal antibody that binds to the PSI domain of the integrin β2 subunit does not block ligand binding, but monoclonal antibodies to the adjacent, putative, I-like domain do (Ref. 17, and C. Huang and T. A. Springer, unpublished). MET (Ref. 18) and some semaphorins[19,20] dimerize, and the PSI domain might be involved in this process. Plexins can also bind to plexin molecules on other cells and, thereby, mediate cell adhesion through a homophilic binding mechanism[9].

In summary, currently, the PSI domain can be described, only vaguely, as an extracellular, putative, protein-binding domain, but its detection in this family of proteins should enable structural studies that provide further insights into the functions of the proteins shown in Fig. 1 (see also Box 1).

Box 1. Note added in proof

After submission of this manuscript, Winberg *et al.*[22] reclassified semaphorins by including plexins, VESPR and MET, in which they also noted G-P motifs (part of the IPT domain). Furthermore, a similarity between some transcription factors and plexin has been noted recently[23].

Acknowledgements

P. B., B. S. and T. D. are supported by the EU and the DFG; T. A. S. is supported by the NIH.

PEER BORK AND TOBIAS DOERKS

EMBL, Meyerhofstr. 1, 69012 Heidelberg, Germany; and Max-Delbrück-Centrum, Robert-Roessle-Strasse 10, 13125 Berlin, Germany.

TIMOTHY A. SPRINGER

Center for Blood Research, Harvard Medical School, Boston, MA 02115, USA.

BEREND SNEL

EMBL, Meyerhofstr. 1, 69012 Heidelberg, Germany.

References

1 Hynes, R. O. (1992) *Cell* 69, 11–25
2 Lee, J-O., Rieu, P., Arnaout, M. A. and Liddington, R. (1995) *Cell* 80, 631–638
3 Tozer, E. C. *et al.* (1996) *J. Biol. Chem.* 271, 21978–21984
4 Tuckwell, D. S. and Humphries, M. J. (1997) *FEBS Lett.* 400, 297–303
5 Tamkun, J. W. *et al.* (1986) *Cell* 46, 271–282
6 Calvete, J. J., Henschen, A. and González-Rodríguez, J. (1991) *Biochem. J.* 274, 63–71
7 Altschul, S. F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
8 Kamayama, T. *et al.* (1996) *Biochem. Biophys. Res. Commun.* 226, 396–402
9 Ohta, K. *et al.* (1995) *Neuron* 14, 1189–1199
10 Kolodkin, A. L., Matthes, D. J. and Goodman, C. S. (1993) *Cell* 75, 1389–1399
11 Nagle, D. L. *et al.* (1999) *Nature* 398, 148–152
12 Comeau, M. R. *et al.* (1998) *Immunity* 8, 473–482
13 Birney, E., Thompson, J. D. and Gibson, T. J. (1996) *Nucleic Acids Res.* 24, 2730–2739
14 Rost, B., Sander, C. and Schneider, R. (1994) *CABIOS* 10, 53–60
15 Eddy, S. R. (1998) *Bioinformatics* 14, 755–763
16 Zhou, L. *et al.* (1997) *Mol. Cell. Neurosci.* 9, 26–41
17 Huang, C., Lu, C. and Springer, T. A. (1997) *Proc. Natl. Acad. Sci. U. S. A.* 94, 3156–3161
18 Gaudino, G. A. *et al.* (1994) *EMBO J.* 13, 3524–3532
19 Koppel, A. M. and Raper, J. A. (1998) *J. Biol. Chem.* 273, 15708–15713
20 Klostermann, A. *et al.* (1998) *J. Biol. Chem.* 273, 7326–7331
21 Bork, P. and Bairoch, A. (1995) *Trends Biochem. Sci.* 20, poster C02
22 Winberg, M. L. *et al.* (1998) *Cell* 95, 903–916
23 Aravind, L., Dixit, V. M. and Koonin, E. V. (1999) *Trends Biochem. Sci.* 24, 44–53