Insulin and Epidermal Growth Factor Receptors Contain the Cysteine Repeat Motif Found in the Tumor Necrosis Factor Receptor

C.W. Ward,[1] P.A. Hoyne,[1] and R.H. Flegg[2]
[1]CSIRO, Division of Biomolecular Engineering, Parkville, Victoria, Australia 3052, and [2]Australian National Sequence Analysis Facility, Walter and Eliza Hall Institute, Parkville, Victoria, Australia 3050

ABSTRACT **The insulin receptor (INSR) and epidermal growth factor receptor (EGFR) are representatives of two structurally related subfamilies of tyrosine kinase receptors. Using the Wisconsin GCG sequence analysis programs, we have demonstrated that the cysteine-rich regions of INSR and EGFR conform to the structural motif found in the tumor necrosis factor receptor (TNFR) family. The study also revealed that these regions were not composed of simple repeats of eight cysteine residues as previously proposed and that the second Cys-rich region of EGFR contained one fewer TNFR repeat than the first. The sequence alignments identified two cysteine residues in INSR that could be responsible for the additional disulfide bonds known to be involved in dimer formation. The published data on the alignments for the fibronectin type III repeat region of the INSR together with previous cysteine mutagenesis studies indicated that there were two disulfide bonds linking the α and β chains of the INSR, but only one α–β linkage in the insulin-like growth factor 1 receptor (IG1R). Database searches and sequence alignments showed that the TNFR motif is also found in the cysteine-rich repeats of laminins and the noncatalytic domains of furin-like proteases. If the starting position of the repeat is altered the characteristic laminin repeat of eight cysteine residues can be shown to consist of a TNFR-like motif fused to the last half of an EGF-like repeat. The overlapping regions of these two motifs are known to have identical disulfide bonding patterns and similar protein folds.**

Key words: cysteine-rich domains, disulfide bond predictions, laminins, profile searching, protein structure, sequence analysis

INTRODUCTION

The insulin receptor (INSR) and epidermal growth factor receptor (EGFR) are representatives of two structurally related subfamilies of tyrosine kinase receptors. The tyrosine kinase receptors are modular proteins[1] that contain highly conserved cytoplasmic kinase domains, flanked by divergent noncatalytic regulatory regions. The cytoplasmic domains are connected to the large extracellular ligand-binding domains by a single transmembrane sequence.[2] The ectodomains of the INSR and EGFR families are more closely related to each other than to the platelet-derived growth factor receptor (PDGFR), fibroblast growth factor receptor (FGFR), vascular endothelial growth factor receptor,[3] Eph/Elk receptor,[4] *axl* receptor,[5] and *tie* tyrosine kinase receptor families.[6] These latter receptors contain variable numbers of immunoglobulin domains and in some cases two (*axl*, Eph/Elk) or three (*tie*) fibronectin type III (Fn3) repeats and distinct Cys-rich regions (Eph/Elk, *tie*).

The ectodomain of the INSR has been shown to contain two homologous domains (L1 and L2) separated by a single Cys-rich region (residues 155 to 312) containing twenty-six Cys residues predicted to be arranged as three repeats of eight.[7] The C-terminal portion of the INSR ectodomain (residues 594 to 929) is comprised of two Fn3 repeats, the first of which contains an insert domain that includes the α–β cleavage site and the alternatively spliced exon 11.[5,8] The EGFR ectodomain was shown to have a similar modular arrangement of L1, Cys-rich and L2 domains but with a second Cys-rich region (residues 475 to 612, containing 21 Cys residues) replacing the much larger unrelated sequence in the INSR[7] (see Fig. 7).

Apart from these predictions very little is known about the secondary and tertiary structure of the ectodomains of these receptor families or their structural relationships to other Cys-rich proteins. The recently reported 3D structure[9] for the ectodomain (residues 1–182) of the tumor necrosis factor receptor-I (TNFR-I) prompted us to examine whether the novel fold seen in this Cys-rich protein was also

Received September 23, 1994; revision accepted January 24, 1995.

Address reprint requests to Dr. C. W. Ward, CSIRO, Division of Biomolecular Engineering, Parkville, Victoria Australia, 3052.

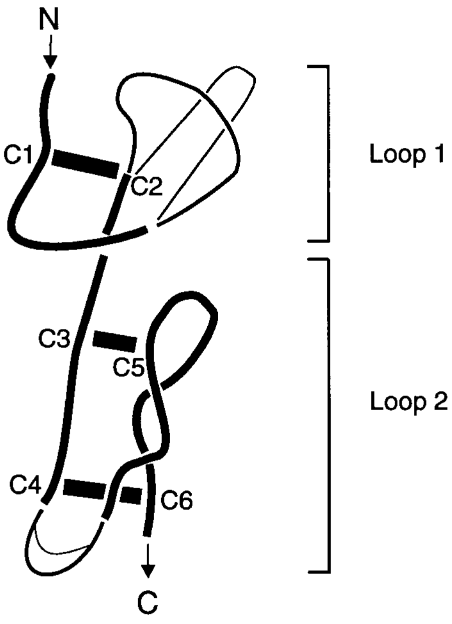


Fig. 1. Schematic representation of the TNFR domain fold corresponding to a single cysteine repeat motif.[9,12] It can be visualized as a double loop structure held together by three disulfide bridges (C1–C2, C3–C5, and C4–C6; very thick lines), where occasionally the C3–C5 disulfide bond is missing. Some members of the TNFR family contain truncated repeats in which either the loop 1 or loop 2 sequences are absent.[9,12] The regions of the structure that are conserved between different repeat domains of TNFR are shown by medium thick lines and different options for the nonconserved regions, by thin lines.[9]

present in the INSR, EGFR and other proteins outside the TNFR family. The TNFR ectodomain is similar in size and composition to the Cys-rich regions of the INSR and EGFR proteins, and the INSR[8,10] resembles the TNFR family[11] in chromatographing as an elongated, asymmetrical molecule on gel filtration.

The 3D structure reveals that TNFR exists as an elongated end-to-end assembly of four similar folding domains each containing six cysteine residues in a double loop motif of approximately 40 amino acids[9] (Fig. 1). Each repeat is held together by conserved C1–C2 (loop 1), C3–C5, and C4–C6 (loop 2) disulfide linkages with the C3–C5 bond occasionally missing.[9,12,13] Sequence comparisons between other members of the TNFR family show that exact half repeats of approximately 20 residues are missing in some of these proteins (see Fig. 2) and that the resulting half repeats are likely to contain an intact C1–C2 or C4–C6 loop.[12]

In this report we use sequence profile searches[14,15] to demonstrate that the Cys-rich regions of the INSR and EGFR contain repeats of the structural motif found in the TNFR family. Disulfide bond assignments and predictions of the total disulfide bonding pattern of the INSR and EGFR families have been made on the basis of these sequence analyses and enzymic digestion[16–20] and cysteine mutagenesis studies.[21,22] In addition, database searches and sequence alignments indicate that the TNFR motif is even more widespread and is found in other proteins such as laminins, furin-like protease precursors and protein J5 from vaccinia virus.

MATERIALS AND METHODS

Protein Sequences

Protein sequences were obtained from version 28 of the SwissProt database[23] which contains 36,018 entries. The Scrutineer program[24] was used to generate a Cys-rich subset (referred to here as SwissC; 6,428 entries) of SwissProt entries greater than 100 residues in length. A sequence was defined as Cys-rich if it contained a stretch of 40 amino acids with a minimum of four cysteines.

Fifty-one sequences were identified in the SwissProt database as belonging to the TNFR family (21 sequences), INSR family (19 sequences), or EGFR family (11 sequences), on the basis of text contained within the comment or keyword field. Forty-nine of these entries were present in SwissC. For completeness the two missing sequences (INSR_DROME and KROS_HUMAN) were added to SwissC. Ten of the 19 INSR sequences were ignored in subsequent analyses because they either lacked (INSR_DROME, KLTK_HUMAN, KLTK_MOUSE, KROS_AVISU, KROS_CHICK, KROS_HUMAN), or contained distorted fragments (TRKA_HUMAN, TRKB_MOUSE, TRKC_PIG, TRKC_RAT) of the INSR Cys-rich region.

Multiple Sequence Analysis

The above sequences were analyzed using the software package of the Genetics Computer Group (Program manual for the GCG package, version 7, April 1991, 575 Science Drive, Madison, WI, USA[14,15]). Repeat sequences in individual proteins were identified using Compare and DotPlot. Files of the Cys-rich repeats of individual proteins were obtained using the Seqed program and aligned using Pileup. Final adjustment to these alignments were made manually if required using Lineup. Profiles were generated from these aligned sequences using ProfileMake. Specific sequences were analysed using Gap or ProfileGap. Databases were probed using ProfileSearch and the alignments were displayed using ProfileSegments. In all cases the default symbol comparison tables were used. ProfileSearch jobs were run with the option/nonormalize since the profiles were short and were being used to identify short repeats in large multidomain proteins.[25]

RESULTS

Profile Generation

The SwissProt database contained 21 sequences from 11 distinct members of the TNFR superfamily

```
                     C1                  C2  C3   C4              C5          C6
Repeat 1
INSR_HUMAN      191  VCPTICKSHGCTAEGL... C.. CHSE CLG..N........ CSQPDDPTKCV 226
EGFR_HUMAN-I    190  ICAQQCSGR.CRGKSPSD. C.. CHNQ CAA..G........ CTGPRE.SDCL 225
EGFR_HUMAN-II   514  KCKLLEGEPREFVEN.SE. CIQ CHPE CLP..QAMN.I.T. CTGRG.PDNCI 556
41BB_MOUSE       28  ................... ... C.DN CQP..GTF...... CRKYN..PVCK  45
CD27_HUMAN       26  SCPE..RHYWAQ.GKL... C.. C.QM CEP..GTFLVK.D. CDQHRKAAQCD  63
CD30_HUMAN       28  TCHGNPSHYYDKAVRR... C.. C.YR CPM..GLFPTQ.Q. CPQ.R.PTDCR  66
CD40_MOUSE       25  TCSD..KQYLHD.GQ.... C.. C.DL CQP..GSRLTS.H. CTALEK.TQCH  60
FASA_HUMAN       59  ................... ... CHKP CPP..GERKAR.D. CTVNGDEPDCV  83
NGFR_HUMAN       31  ACPT..GLYTH.SGE.... C.. C.KA CNL..GEGVAQ.P. CGAN.Q.TVCE  65
OX40_RAT         25  NC.V..KDTY.PSGHK... C.. C.RE CQP..GHGMVS.R. CDHTR.DTVCH  60
TNR1_HUMAN       43  VCPQ..GKYIHPQNNSI.. C.. C.TK CHK..GTYLYN.D. CPGPGQDTDCR  82
TNR2_HUMAN       39  TCRL..REYYDQTAQM... C.. C.SK CSP..GQHAKV.F. CTKTS.DTVCD  76
Profile 1         1  TCPDNPKHYFHQSGQMI.. C.. C.KK CQP..GTWLTQ.D. CTQTRQDTVCH  42

Repeat 2
INSR_HUMAN      240  TCPP.PYYHFQDWR..... CVN .FSF CQDLHHKC(7)CH. QYVI.HNNKCI 285
EGFR_HUMAN-I    239  TCPPLMLYNP(13)SFGAT CV. ..KK CPR.NYVV...... .TD...HGSCV 284
EGFR_HUMAN-II   570  TCPAGVMGENN(7)DAGHV CHL CHPN CT..YG.......  C.TGPGLEGCP 613
41BB_MOUSE       46  SCPP.STF.SSIGG..QPN CNI C.RV C...AGYFRFKKF. CSSTH.NAEC.  85
CD27_HUMAN       64  PCIP.GVSF.SPDHHTRPH CES C.RH CN..SGLL.VR.N. CTITA.NAEC. 105
CD30_HUMAN       68  QCEP.DYYLDEA....DR. CTA C.VT CSRD.DLVEKT.P. CAWNS.SRVC. 106
CD40_MOUSE       61  PCDS.GEF.SAQWNRE.IR CHQ H.RH CEPNQGLRVKK.E. GTAES.DTVC. 103
FASA_HUMAN       84  PCQE.GKEYT.DKAHFSSK CRR C.RL CDEGHGLEVEI.N. CTRTQ.NTKC. 127
NGFR_HUMAN       66  PCLDSVTF.SDVVSAT.EP CKP C.TE C...VGLQSMSAP. CVEAD.DAVC. 107
OX40_RAT         61  PCEP.GFY.NEAVN..YDT CKQ C.TQ CNHRSGSELKQ.N. CTPTE.DTVC. 102
TNR1_HUMAN       83  ECES.GSFTASENH..LRH CLS C.SK CRKEMGQVEIS.S. CTVDR.DTVC. 125
TNR2_HUMAN       77  SCED.STYTQLWNWV..PE CLS CGSR CS..SDQVETQ.A. CTREQ.NRIC. 118

Repeat 3
INSR_HUMAN      287  ECP.SGYTM(6)CTPCLGP C.. .PKV CHLLEG.EKTIDS. VTSAQELRGCT 334
EGFR_HUMAN-I    286  ACGADSYEM(7)CKKCEGP C.. .RKV CNGIGIGEFKDSLSINATNIKHFKNCT 339
41BB_MOUSE       86  ECI.EGFHCL.GPQ..CTR CE. ..KD CR..PGQELT.... ......KQGCK 117
CD27_HUMAN      106  ACR.NGWQCRDKE...CTE CD. .... .............. ........... 121
CD30_HUMAN      107  ECR.PGMFCSTSAVNSCAR CFF H.SV CP..AGMIVKF.P. GTAQK.NTVCE 150
CD40_MOUSE      104  TCK.EGQHCTSKD...CEA CAQ H.TP CIP..GFGVME.M. ATETT.DTVCH 144
FASA_HUMAN      128  RCK.PNFFCNSTV...CEH CDP C.TK CE..HG..IIK.E. CTLTS.NTKCK 166
NGFR_HUMAN      108  RCA.YGYYQDET....TGR CEA C.RV CEAGSGL.V.F.S. CQDKQ.NTVCE 147
OX40_RAT        103  .................. C.. ...Q CRP..GTQPRQ.D. SSHKL.GVDCV 123
TNR1_HUMAN      126  GCR.KNQYRHYWSEN.LFQ CFN C.SL CL..NGT.VHL.S. CQEKQ.NTVC. 166
TNR2_HUMAN      119  TCR.PGWYCALSKQEGCRL CAP L.RK CRP..GFGVAR.P. GTETS.DVVCK 162

Repeat 4
41BB_MOUSE      118  TCS.LGTF.NDQNG..TGV CRP W.TN CSL.DGRSVLK.T. GTTEK.DVVCG 159
CD30_HUMAN      243  QCE.PDYYLDE.....AGR CTA C.VS CSRDDLVE.KT.P. CAWNS.SRTCE 282
CD40_MOUSE      145  PCP.VGFF.SNQSSL.FEK CYP W.TS CED.KNLEVLQ.K. GTSQT.NVICG 187
NGFR_HUMAN      148  ECP.DGTY.SDEANH.VDP CLP C.TV CE..DTERQLR.E. CTRWA.DAECE 189
OX40_RAT        124  PCP.PGHF.SPGS...NQA CKP W.TN CTL.SGKQIRH.P. ASNSL.DTVCE 164
TNR1_HUMAN      167  TCH.AGFFL.RENE..... CVS C.SN CK..KS...L..E. CTK.....LCL 196
TNR2_HUMAN      163  PCA.PGTF.SNTTSS.TDI CRP H.QI CN....V.VAI.P. GNASR.DAVCT 201
Profile 2         1  PCEDPGFYYSDENNHTCER CEP CKTN CEPDSGLEVMKKP. CTETSDDTVCE  50
```

Fig. 2. Alignment of TNFR repeat motifs in members of the TNFR, INSR, and EGFR families. Profile 1 (repeat 1) and profile 2 (repeats 2, 3 and 4) were generated from the aligned TNFR family sequences as described in Methods. The INSR_HUMAN and EGFR_HUMAN sequences were not used for profile generation. The cysteine residues are blocked. The disulfide bond arrangements for the six cysteines in the TNFR motif as well as the internal loop cysteines are indicated. I and II denote the first and second Cys-rich regions of EGFR. All sequences are denoted by their SwissProt accession codes.

TABLE I. Effect of Gap Penalties on the Relative Ranking of TNFR, INSR, and EGFR Sequence Alignments*

	Ranking					Ranking			
	3.0	3.0	2.0	2.0[†]		3.0	3.0	2.0	2.0[†]
Sequence	0.05	0.3	0.5	1.0[‡]	Sequence	0.05	0.3	0.5	1.0[‡]
TNFR family					INSR family				
41BB_MOUSE	10	12	16	22	7LES_DROME	>500	723	790	749
CD27_HUMAN	17	6	6	8	7LES_DROVI	181	268	178	179
CD30_HUMAN	128	7	7	6	IG1R_HUMAN	51	172	157	44
CD40_HUMAN	6	9	11	11	IG1R_RAT	381	182	167	67
CD40_MOUSE	2	5	5	5	INSR_HUMAN	486	59	30	19
FASA_HUMAN	4	4	4	4	INSR_MOUSE	238	36	24	16
FASA_MOUSE	5	3	3	3	INSR_RAT	279	45	26	17
NGFR_CHICK	221	11	10	9	IRR_CAVPO	338	22	15	15
NGFR_HUMAN	53	16	13	14	IRR_HUMAN	231	60	28	21
NGFR_RAT	71	14	12	13					
OX40_RAT	1	1	1	1	EGFR family				
TNR1_HUMAN	3	2	2	2	EGFR_CHICK	262	141	81	85
TNR1_MOUSE	22	8	8	7	EGFR_DROME	86	23	67	207
TNR1_RAT	9	10	9	10	EGFR_HUMAN	354	202	281	111
TNR2_HUMAN	414	17	14	12	ERB2_HUMAN	253	147	79	62
TNR2_MOUSE	>500	196	159	165	ERB3_HUMAN	363	123	121	101
VA53_VACCC	>500	1363	746	445	KER1_CHICK	>500	72	39	73
VA53_VACCV	>500	1474	803	483	KER2_CHICK	>500	54	31	57
VC22_VARV			100	39	KERB_AVIER	>500	62	36	63
VT2_MYXVL	>500	430	378	123	LT23_CAEEL	15	13	20	35
VT2_SFVKA	452	56	51	24	NEU_RAT	494	126	93	141
					XMRK_XIPMA	216	105	73	45

*Rankings result from searching the SwissC database (see Methods) with TNFR profile 2.

[†] Gap weight.

[‡] Gap length weight.

(Table I). Sequences representing nine of the distinct members were used for profile generation and the alignments obtained are shown in Figure 2. The poxvirus sequences and the homologues for CD40, FASA, NGFR, TNR1, and TNR2 from other species were not included to avoid bias. The sequences were aligned using Pileup with gap and gap length penalties of 1 and 0.1 respectively and finally adjusted by eye. The default parameters of 3.0 (gap) and 0.1 (gap length) give poorer alignments of the conserved Cys residues known to be homologous from the 3D structure.[9] The start and stop positions for the repeat regions of the members of the TNFR family are taken from the 3D structure of TNFR-1[9] and previous alignments[13] and correspond to the boundaries between each of the discrete domains.[9] Profile 1 was generated from the seven TNFR family repeat 1 sequences that contain the adjacent cysteine residues at the C2–C3 positions (Fig. 2). The 41BB_MOUSE and FASA_HUMAN repeat 1 sequences were excluded as they lack the first half repeat that contains the C1 and C2 residues. Profile 2 was constructed from the two remaining repeat 1 sequences (41BB_MOUSE and FASA_HUMAN) and the 25 TNFR family repeat 2, 3, and 4 sequences (Fig. 2). The INSR_HUMAN and EGFR_HUMAN sequences were not used for profile generation.

Database Searching

Preliminary analyses revealed that the individual TNFR domain sequences and the TNFR profiles aligned with sequences in the INSR and EGFR families. To establish whether this was a genuine relationship or merely a reflection of cysteine composition, the 6,430 proteins in the SwissC subset of SwissProt were searched with profile 2 and a series of gap/gap length penalties.

The structural constraints of the disulfide bonding pattern that defines the TNFR double loop[9] imply that alignments incorporating similar sequences should not contain either too many gaps or gaps of extreme length within either of the loop motifs. The latter problem arose when the default penalties of 4.5 (gap) and 0.05 (gap length) were used with the ProfileSearch program. Many of the entries with high Quality scores displayed alignments that exceeded a length of 80 residues, clearly an inaccurate representation of a motif that spans only 50 residues. The database searches were repeated with slightly decreasing values for the gap penalty (as gaps were present in the alignments used to generate the profile) but harshly increasing values for the gap length penalty. The series of gap/gap length penalty combinations employed were 3.0/0.05, 3.0/0.3, 2.0/0.5, and 2.0/1.0.

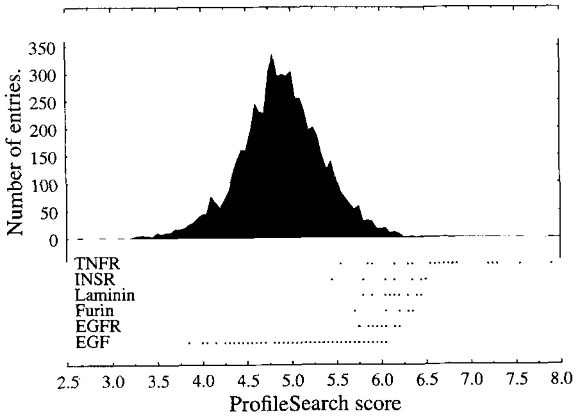


Fig. 3. Distribution of alignment scores of members of the TNFR, INSR, EGFR, laminin, and furin-like protease families compared to all sequences in the SwissC database. The SwissC database (see Methods) was searched with profile 2 at gap/gap length penalties of 2.0 and 1.0, respectively. The dots plotted beside each name indicate the occurrence of scores within a range of 0.05 and not the number of such entries within that range.

The relative rankings for the members of the TNFR, INSR and EGFR families following the four profile searches are summarised in Table I. The results show that the penalty combination of 3.0/0.05 (which is close to the program's default settings of 4.5/0.05) failed to rank TNR2_MOUSE and three of the viral TNFR family in the top 500 even though the alignment of cysteine residues were good and occurred with minimal gaps. In addition most of the INSR and EGFR sequences are ranked outside the top 200 alignments (Table I). The low gap length penalty allowed other sequences to score more highly through the inclusion of substantial gaps. As the gap length penalty was increased the alignments of these other sequences shortened, the number of mismatches increased and their Quality scores dropped.

At gap penalties of 2.0/1.0, the TNFR sequences were clustered at the top of the list with members of the INSR family the next major group (Table I, Fig. 3). The EGFR sequences were less well ranked relative to other proteins but seven appeared within the first 100 sequences (Table I, Fig. 3). Of most interest among the other sequences were the laminins and furin-like proteases. The Cys-rich domains of laminin are classified in SwissProt as EGF-like yet they ranked highly against the TNFR profile (Fig. 3). The Cys-rich regions in the noncatalytic domains of the furin-like proteases are not classified in SwissProt but were found to resemble the laminin repeat and to score highly against the TNFR profile (Fig. 3).

TNFR Repeats in the Insulin Receptor Family

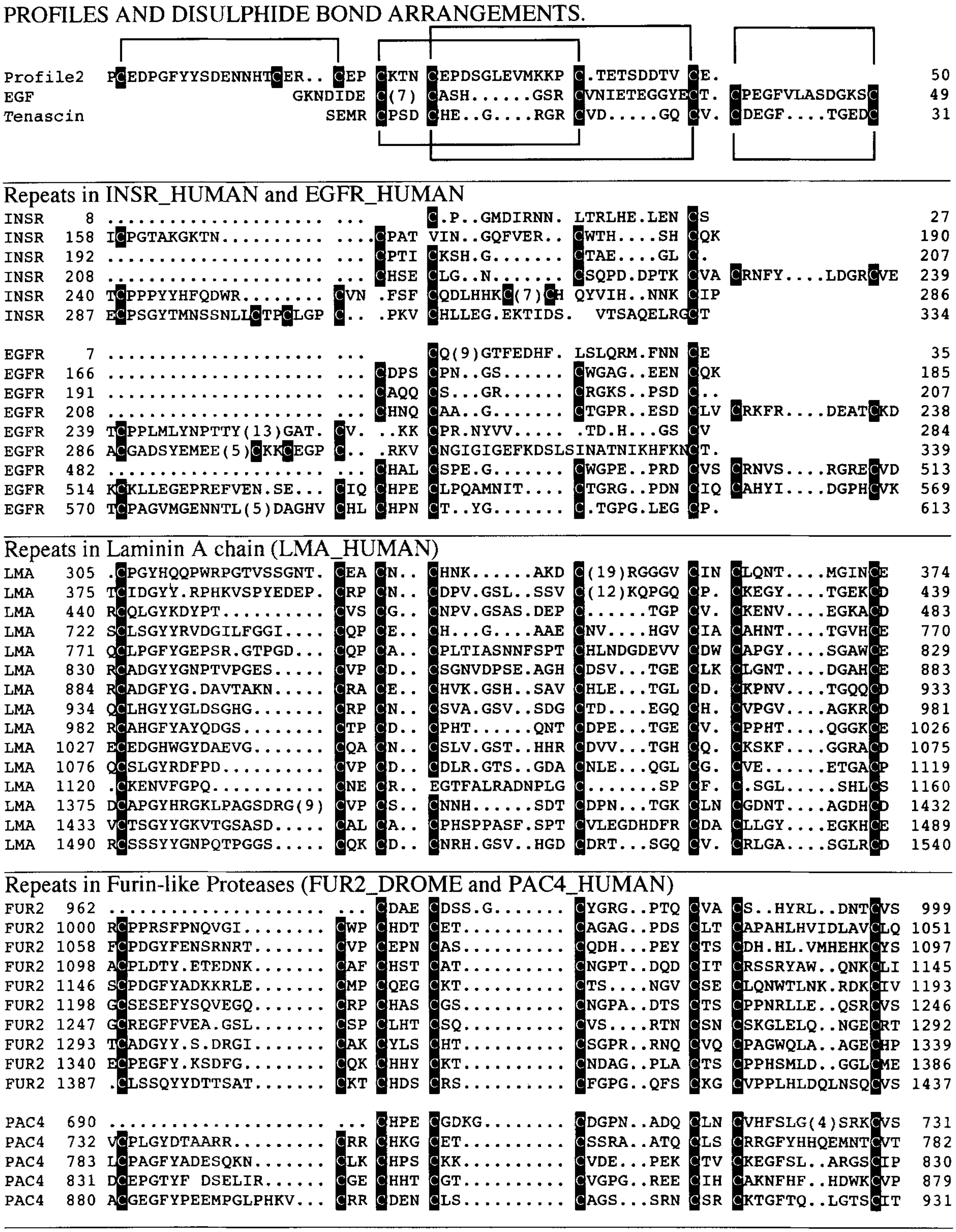
The insulin receptor (INSR_HUMAN), insulin-like growth factor-1 receptor (IG1R_HUMAN), and insulin related receptor (IRR_HUMAN) sequences were searched for additional repeats of the TNFR motif. The penalty combination used was 3.0/0.3 and all three sequences gave similar alignments. Thus only the findings for the INSR_HUMAN will be presented as shown in Figures 2 and 4. The highest scoring repeat identified by Profile 2 was residues 287–334 (Repeat 3 in Fig. 2) which resembles repeat 3 of several of the TNFR proteins in having an extra pair of Cys residues in loop 1 and lacking Cys residues at the C3 and C5 positions (Fig. 2). This repeat extended from the end of the Cys-rich region of INSR_HUMAN to include the pair of cysteines at the start of the L2 domain.[7] The corresponding sequence at the start of the L1 domain (see Fig. 5) matched the second half of profile 2 and aligned Cys residues 8 and 26 with the C4 and C6 positions as expected (Fig. 4).

The next repeat in INSR_HUMAN identified by profile 2 was at residues 240–285. This is equivalent to repeat 2 of the TNFR family. Again there are no Cys residues at the C3 and C5 positions in this domain of the INSR family as is the case in several of the TNFR family repeat sequences (Fig. 2). The alignment also suggested that the Cys residues at 266 and 274 in INSR_HUMAN form an internal disulfide bond in loop 2 presumably to stabilize the extended conformation of this loop expected from the presence of five to eight additional amino acid residues in this domain. The EGFR and the other members of the INSR family lack this insert sequence and the extra pair of cysteines (Fig. 5). In the 3D structure of TNFR-I there is variability between different domains in the conformation of this bottom portion of loop 2.[9]

The INSR family contained another repeat in the region 191–239 that, like repeat 1 of the TNFR family members, contained an adjacent pair of Cys residues at the C2–C3 position (Fig. 2). However the length of this sequence and the presence of 10 Cys residues complicated the profile alignments and there were discrepancies between the results obtained with INSR_HUMAN, IG1R_HUMAN and IR-R_HUMAN. The profile 2 alignment indicated that the repeat covered the whole region equivalent to 191–239 in INSR_HUMAN (Figs. 2 and 4), while the alignment with profile 1 suggested that the repeat stopped short at residue 226. When the sequence 191–226 was examined as two separate sections the alignments revealed that residues 191–207 and 208–226 both matched the loop 2 motif with their four cysteine residues in the C3, C4, C5 and C6 positions (Fig. 4). Thus this region of INSR family differs from the members of the TNFR family in not containing a loop 1 motif (see Fig. 2) and in having an additional 12 residue sequence between the loop 2 repeat at 208–226 and the full repeat at 240–285 (Fig. 4).

TNFR Repeats in the EGFR Family

The EGFR differs from INSR in having two Cys-rich regions. When profile 2 was used to search the whole ectodomain of EGFR_HUMAN, the align-



PROFILES AND DISULPHIDE BOND ARRANGEMENTS.

```
Profile2    PCEDPGFYYSDENNHTCER.. CEP CKTN CEPDSGLEVMKKP C.TETSDDTV CE.                       50
EGF                         GKNDIDE C(7) CASH......GSR CVNIETEGGYECT. CPEGFVLASDGKSC     49
Tenascin                       SEMR CPSD CHE..G....RGR CVD.....GQ CV. CDEGF....TGEDC     31
```

Repeats in INSR_HUMAN and EGFR_HUMAN

```
INSR     8 ..................... ...      C.P..GMDIRNN. LTRLHE.LEN CS                      27
INSR   158 ICPGTAKGKTN.......... ....CPAT VIN..GQFVER.. CWTH....SH CQK                    190
INSR   192 ..................... ... CPTI CKSH.G....... CTAE....GL C.                     207
INSR   208 ..................... ... CHSE CLG..N....... CSQPD.DPTK CVA CRNFY....LDGRCVE   239
INSR   240 TCPPPYYHFQDWR........ CVN .FSF CQDLHHKC(7)CH QYVIH..NNK CIP                    286
INSR   287 ECPSGYTMNSSNLLCTPCLGP C.. .PKV CHLLEG.EKTIDS.  VTSAQELRGCT                     334

EGFR     7 ..................... ...      CQ(9)GTFEDHF. LSLQRM.FNN CE                      35
EGFR   166 ..................... ... CDPS CPN..GS...... CWGAG..EEN CQK                    185
EGFR   191 ..................... ... CAQQ CS...GR...... CRGKS..PSD C..                    207
EGFR   208 ..................... ... CHNQ CAA..G....... CTGPR..ESD CLV CRKFR....DEATCKD   238
EGFR   239 TCPPLMLYNPTTY(13)GAT. CV. ..KK CPR.NYVV..... .TD.H...GS CV                     284
EGFR   286 ACGADSYEMEE(5)CKKCEGP C.. .RKV CNGIGIGEFKDSLSINATNIKHFKNCT.                    339
EGFR   482 ..................... ... CHAL CSPE.G....... CWGPE..PRD CVS CRNVS....RGRECVD   513
EGFR   514 KCKLLEGEPREFVEN.SE... CIQ CHPE CLPQAMNIT.... CTGRG..PDN CIQ CAHYI....DGPHCVK   569
EGFR   570 TCPAGVMGENNTL(5)DAGHV CHL CHPN CT..YG....... C.TGPG.LEG CP.                    613
```

Repeats in Laminin A chain (LMA_HUMAN)

```
LMA    305 .CPGYHQQPWRPGTVSSGNT. CEA CN.. CHNK......AKD C(19)RGGGV CIN CLQNT....MGINCE    374
LMA    375 TCIDGYY.RPHKVSPYEDEP. CRP CN.. CDPV.GSL..SSV C(12)KQPGQ CP. CKEGY....TGEKCD    439
LMA    440 RCQLGYKDYPT.......... CVS CG.. CNPV.GSAS.DEP C......TGP CV. CKENV....EGKACD    483
LMA    722 SCLSGYYRVDGILFGGI.... CQP CE.. CH...G....AAE CNV....HGV CIA CAHNT....TGVHCE    770
LMA    771 QCLPGFYGEPSR.GTPGD... CQP CA.. CPLTIASNNFSPT CHLNDGDEVV CDW CAPGY....SGAWCE    829
LMA    830 RCADGYYGNPTVPGES..... CVP CD.. CSGNVDPSE.AGH CDSV...TGE CLK CLGNT....DGAHCE    883
LMA    884 RCADGFYG.DAVTAKN..... CRA CE.. CHVK.GSH..SAV CHLE...TGL CD. CKPNV....TGQQCD    933
LMA    934 QCLHGYYGLDSGHG....... CRP CN.. CSVA.GSV..SDG CTD....EGQ CH. CVPGV....AGKRCD    981
LMA    982 RCAHGFYAYQDGS........ CTP CD.. CPHT......QNT CDPE...TGE CV. CPPHT....QGGKCE   1026
LMA   1027 ECEDGHWGYDAEVG....... CQA CN.. CSLV.GST..HHR CDVV...TGH CQ. CKSKF....GGRACD   1075
LMA   1076 QCSLGYRDFPD.......... CVP CD.. CDLR.GTS..GDA CNLE...QGL CG. CVE......ETGACP   1119
LMA   1120 .CKENVFGPQ........... CNE CR.. EGTFALRADNPLG C.......SP CF. C.SGL.....SHLCS   1160
LMA   1375 DCAPGYHRGKLPAGSDRG(9) CVP CS.. CNNH......SDT CDPN...TGK CLN CGDNT....AGDHCD   1432
LMA   1433 VCTSGYYGKVTGSASD..... CAL CA.. CPHSPPASF.SPT CVLEGDHDFR CDA CLLGY....EGKHCE   1489
LMA   1490 RCSSSYYGNPQTPGGS..... CQK CD.. CNRH.GSV..HGD CDRT...SGQ CV. CRLGA....SGLRCD   1540
```

Repeats in Furin-like Proteases (FUR2_DROME and PAC4_HUMAN)

```
FUR2   962 ..................... ... CDAE CDSS.G....... CYGRG..PTQ CVA CS..HYRL..DNTCVS   999
FUR2  1000 RCPPRSFPNQVGI........ CWP CHDT CET.......... CAGAG..PDS CLT CAPAHLHVIDLAVCLQ  1051
FUR2  1058 FCPDGYFENSRNRT....... CVP CEPN CAS.......... CQDH...PEY CTS CDH.HL.VMHEHKCYS  1097
FUR2  1098 ACPLDTY.ETEDNK....... CAF CHST CAT.......... CNGPT..DQD CIT CRSSRYAW..QNKCLI  1145
FUR2  1146 SCPDGFYADKKRLE....... CMP CQEG CKT.......... CTS....NGV CSE CLQNWTLNK.RDKCIV  1193
FUR2  1198 GCSESEFYSQVEGQ....... CRP CHAS CGS.......... CNGPA..DTS CTS CPPNRLLE..QSRCVS  1246
FUR2  1247 GCREGFFVEA.GSL....... CSP CLHT CSQ.......... CVS....RTN CSN CSKGLELQ..NGECRT  1292
FUR2  1293 TCADGYY.S.DRGI....... CAK CYLS CHT.......... CSGPR..RNQ CVQ CPAGWQLA..AGECHP  1339
FUR2  1340 ECPEGFY.KSDFG........ CQK CHHY CKT.......... CNDAG..PLA CTS CPPHSMLD..GGLCME  1386
FUR2  1387 .CLSSQYYDTTSAT....... CKT CHDS CRS.......... CFGPG..QFS CKG CVPPLHLDQLNSQCVS  1437

PAC4   690 ..................... ... CHPE CGDKG........ CDGPN..ADQ CLN CVHFSLG(4)SRKCVS   731
PAC4   732 VCPLGYDTAARR......... CRR CHKG CET.......... CSSRA..ATQ CLS CRRGFYHHQEMNTCVT   782
PAC4   783 LCPAGFYADESQKN....... CLK CHPS CKK.......... CVDE...PEK CTV CKEGFSL..ARGSCIP   830
PAC4   831 DCEPGTYF DSELIR...... CGE CHHT CGT.......... CVGPG..REE CIH CAKNFHF..HDWKCVP   879
PAC4   880 ACGEGFYPEEMPGLPHKV... CRR CDEN CLS.......... CAGS...SRN CSR CKTGFTQ..LGTSCIT   931
```

Fig. 4. Alignment of TNFR (profile 2), tenascin, and EGF profiles with the repeat sequences for INSR_HUMAN, EGFR_HUMAN, the laminin A chain, and the human and *Drosophila melanogaster* furin-like protease precursors. The EGF profile was generated from repeats 2 to 9 in the mouse EGF precursor. The tenascin profile was generated from the 15 repeats in the porcine tenascin sequence. The cysteine residues are blocked. The disulfide bond arrangements for the cysteines in the TNFR, EGF, and tenascin motifs are indicated. The laminin repeat is currently listed in version 28 of SwissProt as starting at the third Cys residue 2 in the above alignments. The repeats in the furin-like proteases are listed as stating between the first and second Cys residues.

ment was spread across residues 474–559 at the start of the second Cys-rich region (Figs. 5 and 6). When the profile search was restricted to residues 1–400, the alignment was spread across residues 169–228, at the start of the first Cys-rich repeat (Figs. 5 and 6).

By restricting the search to residues 190–400, profile 2 identified a repeat at 286–339. This is homologous to residues 287–334, the first repeat found in INSR_HUMAN (Fig. 2) and included an extra pair of Cys residues in loop 1 and lacked Cys residues at the C3 and C5 positions. It also included the pair of Cys residues at the start of the L2 domain. As shown for INSR, the equivalent region of the EGFR L1 domain (residues 7–35) aligned with the second half of profile 2 and placed Cys-7 and Cys-34 at the C4 and C6 positions (Fig. 4).

Additional repeats were identified in EGFR_HUMAN at residues 190–225 and 239–284 (Fig. 2). As found for the INSR family, residues 190–225 of EGFR_HUMAN appeared to comprise two loop 2 motifs as it consisted of two 18 residue halves, each with four cysteines identically spaced at the C3, C4, C5, and C6 positions (Fig. 4). Finally the alignments with profile 2 revealed that the sequence 166–185, at the start of the first Cys-rich region of EGFR_HUMAN corresponds to a loop 2 half repeat (Fig. 4). Bestfit analysis with the sequence 191–207 against residues 1–183 and 208–621 confirmed the alignments with 166–183 (35% identity) and 208–224 (47% identity) shown in Figure 4 and supported the conclusion that 191–207 adopts a loop 2 conformation.

The second Cys-rich region of EGFR_HUMAN (residues 475–612) is not a simple repeat of the first. It contained 21 cysteine residues compared to 25 and lacked the adjacent pair of cysteines equivalent to residues 207–208. When searched with profile 2 the second Cys-rich region of EGFR_HUMAN was shown to contain a single loop 2 sequence at 482–500 (Fig. 4) and two double loop motifs at 514–556 and 570–613 (Figs. 2 and 4). These assignments indicated that the two Cys-rich domains of EGFR_HUMAN were not equivalent. The second differed from the first in having one less full TNFR-like repeat. It also differed in having an additional 12 residue sequence between the half repeat at 482–499 and the full repeat at 514–556 (Figs. 4 and 5).

These findings prompted a reanalysis of the sequence identity between the two Cys-rich regions of EGFR_HUMAN and the identity between the EGFR and INSR family sequences. The final alignments are shown in Figure 5. They are based on (1) the results of the profile searches discussed above, (2) pileup analysis of three members of the INSR family and three members of the EGFR family, and (3) gap analysis of EGFR_HUMAN residues 1–310 versus 311–621, The homologous Cys residues are summarized in Figure 6.

Additional gap analyses between smaller regions of EGFR_HUMAN at penalty ratios of 1/0.1 confirmed the assignments shown in Figures 2 and 5 by showing that the repeat at residues 570–613 aligned better with 239–284 (44% identity) than with 286–339 (25% identity), while 514–556 aligned better with 190–225 (52% identity) rather than 239–284 (26% identity). Gap analysis of the longer sequences 191–283 and 515–612 confirmed these assignments and showed that the 12 residue sequence 226–237 was equivalent to 557–568 (see Fig. 5). A similar 12 residue sequence is located at 501–512 (Fig. 4) and has no equivalent homologue in the corresponding region of the first Cys-rich region of the EGFR_HUMAN. The arrangement of these equivalent residues is summarized in Figure 6 along with the predicted disulfide bond assignments for both receptors.

Other Proteins With TNFR Motifs

The top 50 profile alignments consisted of 17 known TNFR family members, six INSR family sequences, and two EGFR family sequences. The remaining 25 proteins were examined in detail. Eight were found not to conform to the TNFR pattern of Cys residues (COX1_LEITA, COX1_TRYBB, SCP_RAT, SCP1_MOUSE, UL32_HSVEB, K1M2_SHEEP, TRYM_CANFA, and RRPO_TACV) and to be either Cys rich or lack evidence of obvious repeat sequences; three (MSAP_PLAFM, MSAP_PLAFF, and MSAP-_PLAFC) contained two EGF-like motifs; one (ITB2_BOVIN) contained four integrin repeats; and one (OTNC_CAEEL) contained a single agrin-like sequence.[26] Of the other high scoring alignments, two small proteins from vaccinia virus (VJ05_VARV and VJ05_VACCC) gave good double loop alignments with the TNFR profile and appear to be genuine members of the TNFR superfamily although they are not listed in SwissProt as TNFR-like. Similarly the five laminin proteins (LMA_MOUSE, LMA_HUMAN, LMB2_HUMAN, LMB2_MOUSE, and UNC6_CAEEL) and the three furin-like proteinases (FUR2_DROME, FURI_HUMAN, and PAC4_HUMAN) gave good alignments with the TNFR profile. As shown in Figure 4, if the starting position is changed from that currently listed in SwissProt, the laminin repeat of eight Cys residues appears to consist of an overlap of a TNFR repeat (first six cysteines) and an EGF-like repeat (last six cysteines). The overlap region in both repeats contain four Cys residues, have identical disulfide bond patterns, and have similar protein folds.[9,27] Of even greater surprise was the finding that the Cys-rich-repeats of the noncatalytic domains of the furin-like proteases can also be rearranged to resemble the merged TNFR/EGF repeats found in the laminins. However none of these merged repeats was found with the C3/C5 pairs missing, although the first repeat in each of the furin-like proteins was truncated

```
INSR_HUMAN-I     1  HLYP..GEVC..PGMDIR......NNLTRLHEL.ENCSVIEGHLQILLM.....FKTRPEDFR..DLSF.PKLIMITDYLLLFRVYGLESLKDL.FPNLT   80
IG1R_HUMAN-I     1  SLWPTSGEIC.GPGIDIR......NDYQQLKRL.ENCTVIEGYLHILLI.....SK..AEDYR..SYRF.PKLTVITEYLLLFRVAGLESLGDL.FPNLT   81
 IRR_HUMAN-I     1           C..PSLDIR......SEVAELRQL.ENCSVVEGHLQILLM.....FTATGEDFR..GLSF.PRLTQVTDYLLLFRVYGLESLRDL.FPNLA   73
EGFR_HUMAN-I     1     LEEKKVCQGTSNKLTQLGTFEDHFLSLQRMFNNCEVVLGNLEITYV.......QRNYDL...SF.L.KTIQEVAGYVLI....ALNTVERIPLENLQ   81
ERB2_HUMAN-I     1       STQVCTGTDMKLRLPASPETHLDMLRHLYQGCQVVQGNLELTYL.......PTNASL...SF.L.QDIQEVQGYVLI....AHNQVRQVPLQRLR   79
ERB3_HUMAN-I     1  SEVGNSQAVCPGTLNGLSVTGDAENQYQTLYKLYERCEVVMGNLEIVLT.......GHNADL...SF.L.QWIREVTGYVLV....AMNEFSTLPLPNLR   84
INSR_HUMAN-II  311         .VCHLLEGEK...TIDSVTSA..QEL.RGCTVINGSL.II..NIRGG.NNLAAELE.ANLGL...IEEISGYLKIRRSYAL..VSLSFFRKLR  386
IG1R_HUMAN-II  309         ..C...EEEKKTKTIDSVTSA..QML.QGCTIFKGNL.LI..NIRRG.NNIASELE.NFMGL...IEVVTGYVKIRHSHAL..VSLSFLKNLR  383
 IRR_HUMAN-II  299         .EC.....KVGTKTIDSIQAA..QDL.VGCTHVEGSL.IL..NLRQG.YNLEPQLQ.HSLGL...VETITGFLKIKHSFAL..VSLGFFKNLK  371
EGFR_HUMAN-II  312         .VCNGIGIGE.FKDSLSINATNIKHF.KNCTSISGDLHILPVAFRGDSFTHTPPLDPQELDILKTVKEITGFLLI.QAWPENRTDLHAFENLE  400
ERB2_HUMAN-II  320         .VCYGLGM.EHLREVRAVTSANIQEF.AGCKKIFGSLAFLPESFDGDPASNTAPLQPEQLQVFETLEEITGYLYI.SAWPDSLPDLSVFQNLQ  408
ERB3_HUMAN-II  311         .ACEGTGSGSRFQ...TVDSSNIDGF.VNCTKILGNLDFLITGLNGDPWHKIPALDPEKLNVFRTVREITGYLNI.QSWPPHMHNFSVFSNLT  397

INSR_HUMAN-I    81  VIRGSRLFF.N.YALVIFE....MVH...........LKELGLYNLMNITRGSVRIEKNNELCYLATIDWSRILDSVE....DNYIVLNKDDN.EECGDI  158
IG1R_HUMAN-I    82  VIRGWKLFY.N.YALVIFE....MTN...........LKDIGLYNLRNITRGAIRIEKNADLCYLSTVDWSLILDAVS....NNYIVGNKPPK..ECGDL  158
 IRR_HUMAN-I    74  VIRGTRLFL.G.YALVIFE....MPH...........LRDVALPALGAVLRGAVRVEKNQELCHLSTIDWGLLQPAPG....ANHIVGNKLG..EECADV  150
EGFR_HUMAN-I    82  IIRGNMYYE.NSYALAVLSNYDA..NKT........GLKELPMRNLQEILHGAVRFSNNPALCNVESIQWRDIVSSDF.LSNMSMDFQNHLG...SCQ..  164
ERB2_HUMAN-I    80  IVRGTQLFE.DNYALAVLDNGDPLNNTTPVTGASPGGLRELQLRSLTEILKGGVLIQRNPQLCYQDTILWKDIFHKNNQLA.LTLIDTNRS...RACH..  172
ERB3_HUMAN-I    84  VVRGTQVYD.GKFAIFVMLNYNT..NSSH.......ALRQLRLTQLTEILSGGVYIEKNDKLCHMDTIDWRDIV.RDR....DAEIVVKDNG..RSCP..  165
INSR_HUMAN-II  387  LIRGET.LEIGNYSFYALDNQN.LRQ...........LWDWSKHNL.TITQGKLFFHYNPKLCLSE.IHKMEEVSGTKGRQERNDIALKTNGDQASCE..  469
IG1R_HUMAN-II  384  LILGEEQLE.GNYSFYVLDNQN.LQQ...........LWDWDHRNL.TIKAGKMYFAFNPKLCVSE.IYRMEEVTGTKGRQSKGDINTRNNGERASCE..  466
 IRR_HUMAN-II  372  LIRGDAMVD.GNYTLYVLDNQN.LQQ...........LGSWVAAGL.TIPVGKIYFAFNPRLCLEH.IYRLEEVTGTRGRQNKAEINPRTNGDRAACQ..  454
EGFR_HUMAN-II  401  IIRGRTK.QHGQFSLAVV.SLN.ITS...........L.GL..RSLKEISDGDVIISGNKNLCYANTINWKKLF.GTSGQKTK..IISNR.GE.NSCK..  476
ERB2_HUMAN-II  409  VIRGRILHN.GAYSL.TLQGLG.ISW...........L.GL..RSLRELGSGLALIHHNTHLCFVHTVPWDQLFRNPHQALLH..TA.NR.PE.DECV..  484
ERB3_HUMAN-II  398  TIGGRSLYNRG.FSLLIMKNLN.VTS...........L.GF..RSLKEISAGRIYISANRQLCYHHSLNWTKVLRGPTEERLD..IKHNR.PR.RDCV..  475

INSR_HUMAN-I   159  CPG..TAKGKTNCPATVINGQFVERCWTHS..HCQK............VCPTICKSH.GCTAE..GLC..CHSECL..GN...CSQPDDPTKCVACRNFY  232
IG1R_HUMAN-I   159  CPG..TMEEKPMCEKTTINNEYNYRCWTTN..RCQK............MCPSTCGKR.ACTEN..NEC..CHPECL..GS...CSAPDNDTACVACRHYY  232
 IRR_HUMAN-I   151  CPGVLGAAGEP.CAKTTFSGHTDYRCWTSS..HCQR............VCP..CPHGMACTAR..GEC..CHTECL..GG...CSQPEDPRACVACRHLY  224
EGFR_HUMAN-I   165  ...........KCDPSCPNG....SCWGAGEENCQK...LTKI.....ICAQQCSG..RCRGKSPSDC..CHNQCA..AG...CTGPRE.SDCLVCRKFR  231
ERB2_HUMAN-I   173  ...........PCSPMCKGS....RCWGESSEDCQS...LTRT.....VCAGGC.A..RCKGPLPTDC..CHEQCA..AG...CTGPKH.SDCLACLHFN  238
ERB3_HUMAN-I   166  ...........PCHEVC.KG....RCWGPGSEDCQT...LTKT.....ICAPQCNG..HCFGPNPNQC..CHDECA..GG...CSGPQD.TDCFACRHFN  231
EGFR_HUMAN-II  477  ......ATG.QVCHALCSPE....GCWGPEPRDCVSCRNVSRGRECVDKCKLLEGEP.REFVEN.SECIQCHPECLPQAMNITCTGRGPDN.CIQCAHYI  562
ERB2_HUMAN-II  485  ......GEG.LACHQLCARG....HCWGPGPTQCVNCSQFLRGQECVEECRVLQGLP.REYVNA.RHCLPCHPECQPQNGSVTCFGPEADQ.CVACAHYK  570
ERB3_HUMAN-II  476  ......AEG.KVCDPLCSSG....GCWGPGPGQCLSCRNYSRGGVCVTHCNFLNGEP.REFAHE.AECFSCHPECQPMEGTATCNGSGSDT.CAQCAHFR  561

INSR_HUMAN-I   233  LDGRCVETCPPPYYHFQDWR...............CVN.FSFCQD.LHHKCKNSRRQGCHQYVIHN.NKCIPECPSG.YTMNSSN.L.LCTPCLGPCPK.  310
IG1R_HUMAN-I   233  YAGVCVPACPPNTYRFEGWR...............CVD.RDFCANILSAESSDS..EG...FVIHD.GECMQECPSG.FIRNGSQSM.YCIPCEGPCPKV  308
 IRR_HUMAN-I   225  FQGACLWACPPGTYQYESWR...............CVT.AERCASLHSVPGRAS......TFGIHQ.GSCLAQCPSG.FTRNSS.SI.FCHKCEGLCPKE  298
EGFR_HUMAN-I   232  DEATCKDTCPPLMLYNPTTYQMDVNPEGKYSFGATCVK.K..CPRN...............YVVTDHGSCVRACGADSYEM.EEDGVRKCKKCEGPCRK.  311
ERB2_HUMAN-I   239  HSGICELHCPALVTYNTDTFESMPNPEGRYTFGASCVT.A..CPYN...............YLSTDVGSCTLVCPLHNQEVTAEDGTQRCEKCSKPCAR.  319
ERB3_HUMAN-I   232  DSGACVPRCPQPLVYNKLTFQLEPNPHTKYQYGGVCVA.S..CPHN...............FVV.DQTSCVRACPPDKMEV.DKNGLKMCEPCGGLCPK.  310
EGFR_HUMAN-II  563  DGPHCVKTCPAGVMGENNTL.VWKYADA....GHVCHLCHPNCTYG...............CTGPGLEGCPTNGPKIPS                        621
ERB2_HUMAN-II  571  DPPFCVARCPSGVKPDLSYMPIWKFPDE....EGACQPCPINCTHS...............CVDLDDKGCPAEQRAQRASPLTS                   632
ERB3_HUMAN-II  562  DGPHCVSSCPHGVLGAKGP..IYKYPDV....QNECRPCHENCTQG...............CKGPELQDCLGQT                             614
```

Fig. 5.

(Fig. 4). The alignments shown in Figure 4 also revealed that the arrangement of residues 227–239 in INSR_HUMAN and 226–238, 557–569, and 501–513 in EGFR_HUMAN adjacent to the TNFR motifs, resembled the merged TNFR/EGF motif seen with the laminin and furin sequences.

DISCUSSION

Sequence analyses have revealed that many proteins are composed of a number of different, sometimes repeated, structural and functional units.[1] These modules are most widespread among eukaryotic extracellular proteins and are assembled by insertion or duplication of whole exons.[28] At the sequence level they can be quite variable and may be recognizable only by comparison with specific motifs which in turn become increasingly blurred as more and more sequences become available.[1] These motifs are predicted to have similar folding patterns and mediate similar types of interaction.[13] The majority of known modules are characterized by a high cysteine content although some, like the fibronectin type 3 repeats and immunoglobulin domains, are not.[1]

In this paper we have shown that the structurally related, Cys-rich regions of the INSR and EGFR families contain repeats of the structural motif found in members of the TNFR family (see Figs. 1 and 7). This motif (Cys repeat) of approximately 42 amino acids contains six conserved Cys residues and has been identified previously in the low affinity nerve growth factor receptor (NGFR); the T-cell antigens 4-1BB, CD27, CD30, and OX40; the B-cell antigen CD40; two receptors for tumor necrosis factor (TNFR-1 and TNFR-2); Fas (or APO-1, the apoptosis-mediating, cell-surface antigen); and three pox virus protein sequences VC22, VT2, and VA53. Most of these proteins contain four Cys repeat motifs although VC22, VT2, VA53, 4-1BB, CD27, Fas, and OX40 have fewer and CD30 has more.[12,13]

The alignments presented here indicate that the Cys-rich regions of the INSR and EGFR families possess most of the additional features found in members of the TNFR family.[9,12] These are (1) the frequent occurrence of proline near the C1 residue in loop 1, (2) the conserved tyrosine or phenylalanines four to six residues carboxy-terminal of C1, (3) a pair of adjacent Cys residues in the region equivalent to repeat 1 of the TNFR superfamily, and (4) the absence of cysteine residues at the C3 and C5 positions in some of the repeats (Fig. 2).

The results presented here also reveal that the structural boundaries implied by the original description of the L1, L2, and Cys-rich regions of INSR and EGFR[7] require modification. The Cys-rich regions should include the pair of cysteines currently assigned to the N-terminal region of the L2 domains as they are part of a TNFR repeat motif and are predicted to be linked by a disulfide bond. The equivalent pair of cysteine residues at the start of each L1 domain has been shown to be disulfide linked by chemical analysis of tryptic peptides.[20]

Our findings on the homologous cysteines in the two halves of the EGFR ectodomain and the ectodomain of INSR differ from those of Bajaj et al.[7] at 10 positions. Seven of these discrepancies arise from the differences between the number of TNFR motifs in the two Cys-rich regions of EGFR. As shown in Figures 5 and 6, Cys residues 191, 195, 199, 271, 283, 287, and 302 are not homologous to residues 502, 511, 515, 596 600, 604, or 612, respectively, as reported.[7] The remaining three differences result from (1) the confusion caused in the alignments by the additional pair of cysteines at 266 and 274 in the INSR but not the EGFR family or other members of the INSR family (IRR, IG1R); and (2) the significant difference in the spatial arrangement of INSR residues 159, 169, 182, and 188 compared to 166, 170, 175, and 183 of EGFR. The EGFR sequence at 166–183 matches the loop 2 motif and assigns the four cysteines to the C3, C4, C5, and C6 positions. This region of the INSR family also matched the loop 2 motif but only assigned cysteine residues to the C3, C5, and C6 positions. Thus residues 169 and 182 are predicted to be disulfide bonded in the normal C3–C5 manner with 159 and 188 available to participate in atypical disulfide linkages. A less-favored alternative is that 159 and 169 form a small loop 1 domain, leaving 182 and 188 in the nonlinked C5 and C6 positions.

This difference between the two receptors is most significant and may be related to the fact that the INSR is a homodimer. Schaffer and Ljungqvist[18] have shown that Cys-524 is involved in INSR dimer stabilization. However, additional dimer disulfide bonds exist, since an INSR in which Cys-524 has been converted to alanine retains its dimeric status.[22] At least two more dimer bonds must be present since the ectodomain contains an odd number (37) of cysteine residues and no free cysteines.[29,30] These additional dimer bonds are located upstream of residue 524 since the eight cysteine residues in the Fn3 repeats and the insert regions are contained in the 110 kDa fragment (residues 582–1343) obtained by tryptic digestion of native receptor. This fragment is a monomer[16,20,31] and contains all the α–β disulfide linkages.

The alignments suggest that 159 and 188 may provide the additional disulfide bonds between the two monomers of INSR. The chemical reactivity

Fig. 5. Sequence alignments for the ectodomains of three members of each of the INSR and EGFR families. The alignments are based on Pileup of the six sequences with adjustments to accommodate the results of the profile searches and Gap analyses of selected regions. The cysteine residues are blocked. All sequences are denoted by their SwissProt accession codes. I and II denote the first (residues 1–310) and second (residues 311–621) halves of EGFR family ectodomains.

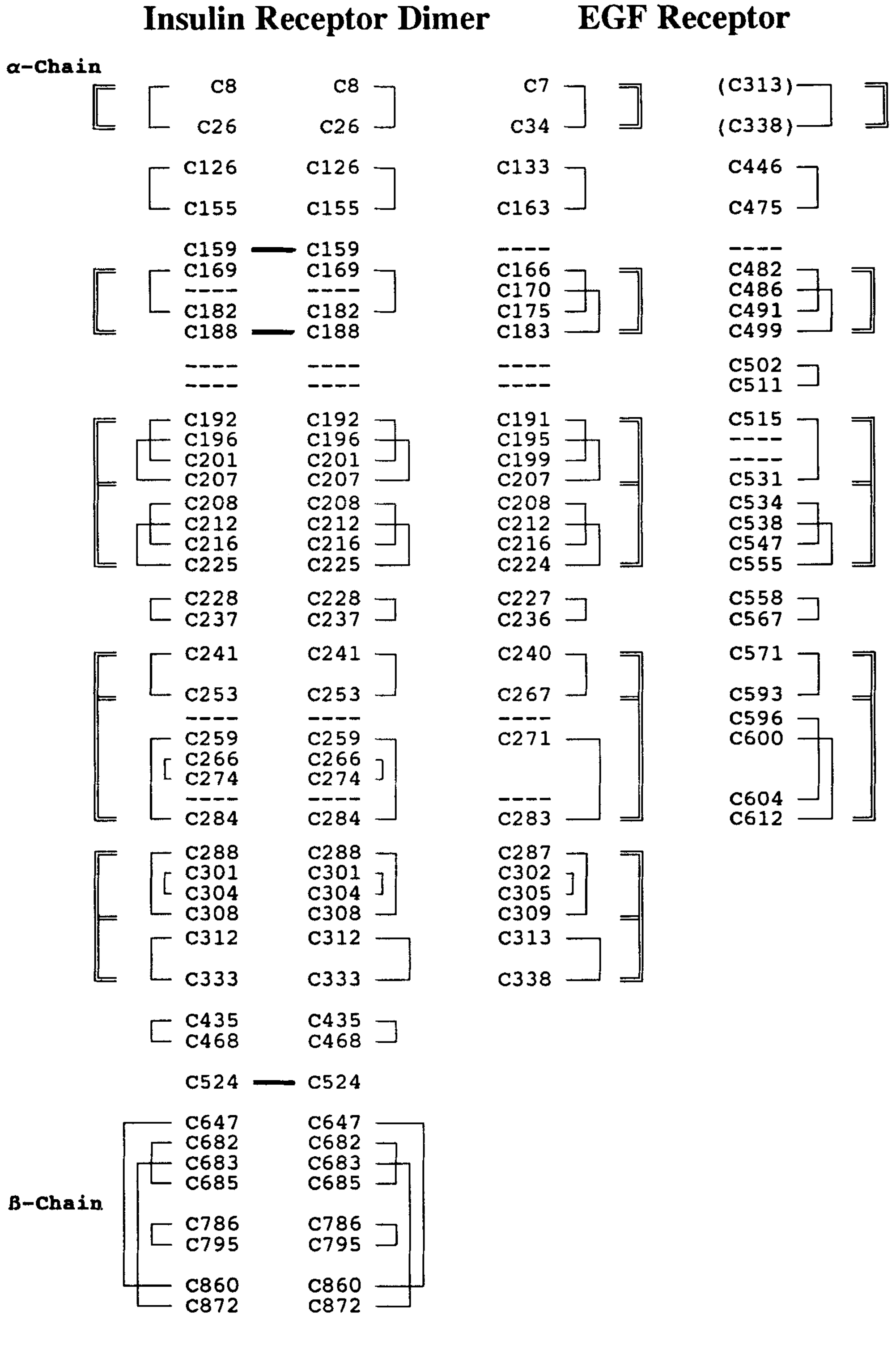


Fig. 6. Summary of equivalent cysteine residues and the disulfide bond arrangements in the ectodomains of INSR_HUMAN and EGFR_HUMAN. The INSR_HUMAN ectodomain is depicted as a dimer with the three dimer disulfide bonds shown with solid bold lines. The residues on each line from INSR_HUMAN and the two halves of EGFR_HUMAN are homologous. The TNFR repeats are enclosed in large vertical double line brackets.

Fig. 7. Organization of the structural domains of INSR, EGFR, and some members of the TNFR family based on the motif representations of Bazan.[12] The Cys repeats in each protein are depicted by the flags; the L1 and L2 domains of INSR_HUMAN and EGFR_HUMAN are represented by the large wavy lines and the Fn3 domains by the overlapping sandwich. The shading of the flags indicates whether that region comprises a full double loop repeat or a truncated single repeat. The small 10 residue loops found in INSR_HUMAN and EGFR_HUMAN are represented by the miniature flags. The tyrosine kinase catalytic domains are represented by the cylinders. The three dimer disulfide bonds in INSR_HUMAN are shown by broken lines.

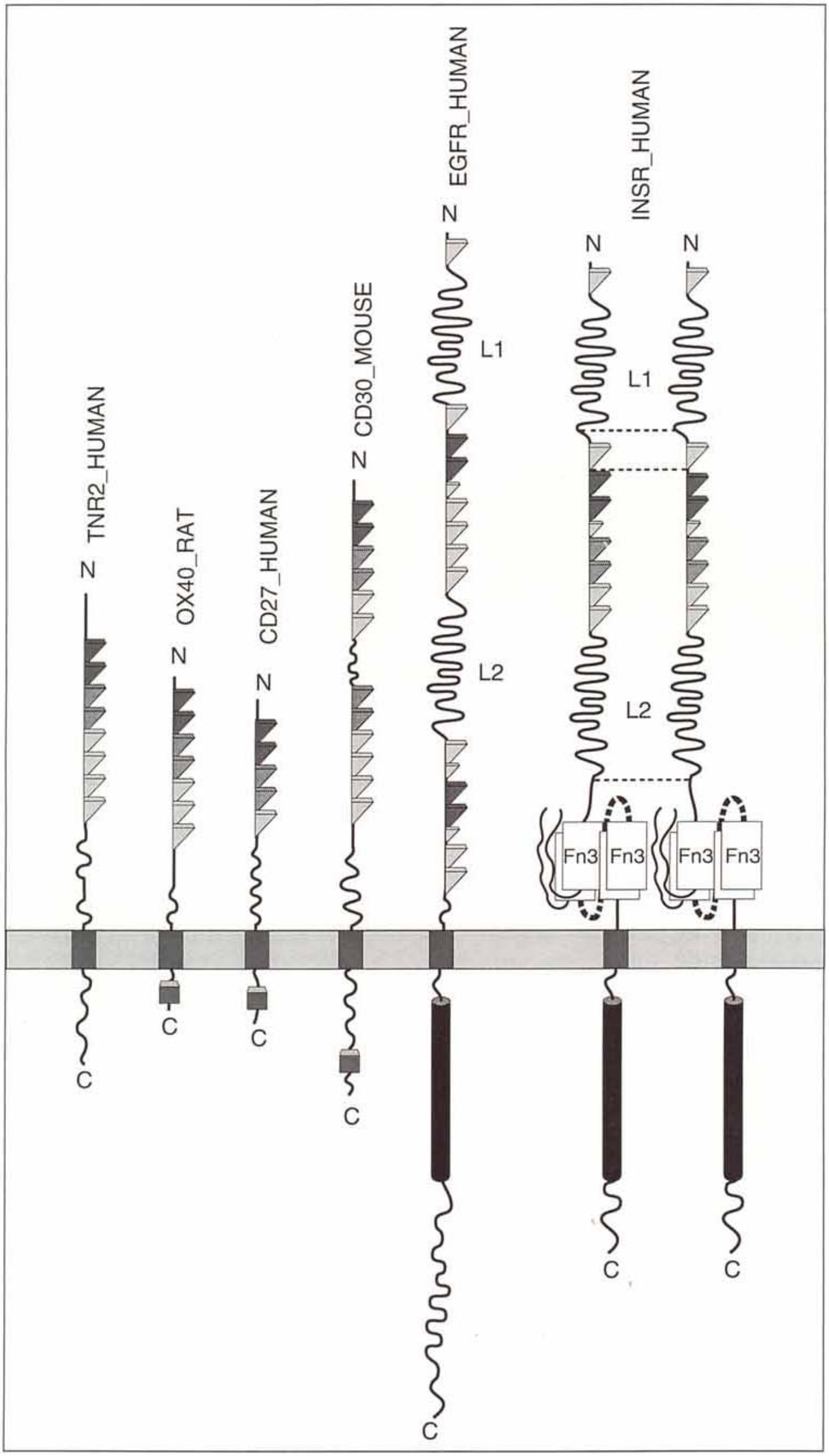

N
TNR2_HUMAN
C
N
OX40_RAT
C
N
CD27_HUMAN
C
N
CD30_MOUSE
C
N
EGFR_HUMAN
L1
L2
C
INSR_HUMAN
N
N
L1
L2
Fn3
Fn3
Fn3
Fn3
C
C

Fig. 7.

data[29,30] indicate that the number of dimer disulfides in INSR is low[22] and the peptide analyses[18,20] support the involvement of 159. The tryptic fragment 1–164 (with five cysteines) was found to be a dimer while peptide 1–122 (with two cysteines) was a monomer.[20] In addition, residues 126 and 155 should be disulfide bonded to match the linkage demonstrated between the equivalent pair, residues 435 and 468, at the C-terminal end of the L2 domain.[18] This leaves 159 as the remaining candidate responsible for the dimeric status of residues 1–164.

Figure 6 also summarizes the options for the disulfide bonding pattern of the C-terminal region of the INSR ectodomain. This region contains the two Fn3 repeats interrupted by a 120 residue insert between residues 648 and 768.[5] Schaefer et al.[8] used these alignments to suggest that the four β-chain residues, 786, 795, 860, and 872, are involved in F–G and C′–E disulfide linkages, respectively, as found in the growth hormone receptor.[32] However, such an arrangement would preclude the existence of the α–β disulfide linkages known to be present. An alternative possibility is shown in Figure 6. It retains the 786–795 intrasheet linkage and suggests that residue 647 in the E strand of the first Fn3 repeat is linked to residue 860 (C′ strand, second Fn3 repeat), leaving 872 (E strand second Fn3 repeat) linked to one of the cysteines in the 682, 683, 685 triplet. The implications of this are that the two Fn3 domains of INSR are expected to be folded back on each other and held together by a complex knot of disulfide bonds. Such a model is facilitated by the fact that there are 10 amino acid residues between the two Fn3 repeats of INSR compared to two in the growth hormone receptor.[32]

The disulfide bond arrangements suggested here result in two α–β disulfide linkages in INSR, as suggested by the chemical reactivity data[29,30] and only one α–β disulfide bond in IG1R. In IG1R the cysteine equivalent to 872 is missing and is replaced by an additional cysteine at the position equivalent to 675 in INSR_HUMAN near the triplet of cysteines in the insert region. The mutagenesis data[21] implicated 647 in an α–β disulfide bond and showed that α–β linkage still occurred in an INSR_HUMAN in which all three cysteines 682, 683, and 685 had been replaced by serine.

The database profile searches indicated that the TNFR motif is even more widespread and is present in other proteins such as laminins and furins. This was intriguing given that the Cys-repeat of laminin is described in SwissProt as an aberrant EGF-like repeat with an extra disulfide link forming a tail loop.[33] As shown in Figure 4, if the start position of the repeat is changed, the eight-cysteine repeats in laminin and furin appear to be chimerics of overlapping TNFR-like (cysteines 1 to 6) and EGF-like (cysteines 4 to 8) motifs.

The proposed relationship between laminin and the TNFR fold has been reported previously. Using the FASTP program, significant homology was found between the Cys-rich region of laminin and OX40[34] and 4-1BB.[35] The latter authors suggested that the ectodomain of 4-1BB contained two laminin-like repeats, each of seven cysteine residues. They also found a functional relationship in their observation that 4-1BB could bind to extracellular matrix components.

CONCLUSION

Previous comparisons of the homology between members of the TNFR and the Cys-rich domains of the low density lipoprotein receptor or EGFR failed to establish significant relationships.[34] However, as Bork[1] points out, the identification of remote relationships by sequence analysis is facilitated by the availability of 3D structural knowledge. In this way the recently reported structure of p55, the type 1 TNFR[9] has allowed previous alignments[13] of members of the TNFR family to be refined and enabled us to extend these comparisons to include members of the INSR, EGFR, laminin, and furin protease families. The latter observations have led to a redefinition of the nature of the laminin repeat and provided a structural definition for the Cys-rich motif in the furin-like proteases.

REFERENCES

1. Bork, P. Mobile modules and motifs. Curr. Biol. 2:413–421, 1992
2. Yarden, Y., Kelman, Z. Transmembrane signalling receptors for cytokines and growth factors. Curr Opinion Str. Biol. 1:582–589, 1991.
3. De Vries, C., Escobedo, J.A., Ueno, H., Houck, K., Ferrara, N., Williams, L.T. The fms-like tyrosine kinase, a receptor for vascular endothelial growth factor. Science 255:989–991, 1992.
4. Lhotak, V., Pawson, T. Biological and biochemical activities of a chimeric epidermal growth factor-Elk receptor tyrosine kinase. Mol. Cell. Biol. 13:7071–7079, 1993.
5. O'Bryan, J.P., Frye, R.A., Cogswell, P.C., Neubauer, A., Kitch, B., Prokop, C., Espinosa, R., III, Le Beau, M.M., Earp, H.S., Liu, E. *axl,* a transforming gene isolated from primary human myeloid leukemia cells, encodes a novel receptor tyrosine kinase. Mol. Cell. Biol. 11:5016–5031, 1991.
6. Sato, T.N., Qin, Y., Kozak, C.A., Audus, K.L. *tie-1* and *tie-2* define another class of putative receptor tyrosine kinase genes expressed in early embryonic vascular system. Proc. Natl. Acad. Sci. U.S.A. 90:9355–9358, 1993.
7. Bajaj, M., Waterfield, M.D., Schlessinger, J., Taylor, W.R., Blundell, T. On the tertiary structure of the extracellular domains of the epidermal growth factor and insulin receptors. Biochim. Biophys. Acta 916:220–226, 1987.
8. Schaefer, E.M., Erickson, H.P., Federwisch, M., Wollmer, A., Ellis, L. Structural organization of the human insulin receptor ectodomain. J. Biol. Chem. 267:23393–23402, 1992.
9. Banner, D.W., D'Arcy, A., Janes, W., Gentz, R., Schoenfeld, H.J., Broger, C., Loetscher, H., Lessiauer, W. Crystal structure of the soluble human 55 kd TNF receptor-human TNFbeta complex: Implications for TNF receptor activation. Cell 73:431–445, 1993.
10. Johnson, J.D., Wong, M.L., Rutter, W.J. Properties of the insulin receptor ectodomain. Proc. Natl. Acad. Sci. U.S.A. 85:7516–7520, 1988.
11. Vissavajjhala, P., Ross, A.H. Purification and characterization of the recombinant extracellular domain of human

nerve growth factor receptor expressed in a baculovirus system. J. Biol. Chem. 265:4746–4752, 1990.

12. Bazan, J.F. Emerging families of cytokines and receptors. Curr. Biol. 3:603–606, 1993.
13. Mallett, S., Barclay, A.N. A new superfamily of cell surface proteins related to the nerve growth factor receptor. Immunol. Today 12:220–223, 1991.
14. Gribskov, M., Luthy, R., Eisenberg, D. Profile analysis. Methods Enzymol. 183:146–159, 1990.
15. Devereux, J., Haeberli, P., Smithies, O. A comprehensive set of sequence analysis programs for the VAX. Nucl. Acids Res. 12:387–395, 1984.
16. Clark, S., Eckardt, G., Siddle, K. Harrison, L.C. Changes in insulin receptor structure associated with trypsin-induced activation of the receptor tyrosine kinase. Biochem. J. 276:27–33, 1991.
17. Kohda, D., Odaka, M., Lax, I., Kawasaki, H., Suzuki, K., Ullrich, A., Schlessinger, J., Inagaki, F. A 40-kDa epidermal growth factor/transforming growth factor alpha-binding domain produced by limited proteolysis of the extracellular domain of the epidermal growth factor receptor J. Biol. Chem. 268:1976–1981, 1993.
18. Schaffer, L., Ljungqvist, L. Identification of a disulfide bridge connecting the alpha-subunits of the extracellular domain of the insulin receptor Biochem. Biophys. Res. Commun. 189:650–653, 1992.
19. Waugh, S.M., DiBella, E.E., Pilch, P.F. Isolation of a proteolytically derived domain of the insulin receptor containing the major site of cross-linking/binding. Biochemistry 28:3448–3458, 1989.
20. Xu, Q-Y., Plaxton, R.J., Fujita-Yamaguchi, Y. Substructural analysis of the insulin receptor by microsequence analyses of limited tryptic fragments isolated by SDS-PAGE in the absence or presence of DTT. J. Biol. Chem. 265:18673–18681, 1990.
21. Cheatham, B., Kahn, C.R. Cysteine 647 in the insulin receptor is required for normal covalent interaction between A- and B-subunits and signal transduction. J. Biol. Chem. 267:7108–7115, 1992.
22. Macaulay, S.L., Polites, M., Hewish, D.R., Ward, C.W. Cysteine-524 is not the only residue involved in the formation of disulfide-bonded dimers of the insulin receptor. Biochem. J. 303:575–581, 1994.
23. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence data bank. Nucl. Acids Res. 19:2247–2249, 1991.
24. Sibbald, P.R., Argos, P. Scrutineer: A computer program that flexibly seeks and describes motifs and profiles in protein sequence databases. CABIOS 6:279–288, 1990.
25. Gibson, T.J., Thompson, J.D., Heringa, J. The KH domain occurs in a diverse set of RNA-binding proteins that include the anti-terminator NusA and is probably involved in binding to nucleic acid. FEBS Lett. 324:361–366.
26. Schwarzbauer, J.E., Spencer, C.S. The *Caenorhabditis elegans* homologue of the extracellular calcium binding protein SPARC/osteonectin affects nematode body morphology and mobility. Mol. Biol. Cell 4:941–952, 1993.
27. Hommel, U., Harvey, T.S., Driscoll, P.C., Campbell, I.D. Human epidermal growth factor: High resolution solution structure and comparison with human transforming growth factor α. J. Mol. Biol. 227:271–282, 1992.
28. Patthy, L. Modular exchange principles in proteins. Curr. Biol. 1:351–361, 1991.
29. Finn, F.M., Ridge, K.D., Hofmann, K. Labile disulfide bonds in human placental insulin receptor. Proc. Natl. Acad. Sci. U.S.A. 87:419–423, 1990.
30. Chiacchia, K.B. Quantitation of the class I disulfides of the insulin receptor. Biochem. Biophys. Res. Commun. 176: 1178–1182, 1991.
31. Shoelson, S.E., White, M.F., Kahn, C.R. Tryptic activation of the insulin receptor. Proteolytic truncation of the alpha subunit releases the beta subunit from inhibitory control. J. Biol. Chem. 263:4852–4860, 1988.
32. De Vos, A.M., Ultsch, M., Kossiakoff, A.A. Human growth hormone and extracellular domain of its receptor: Crystal structure of the complex. Science 255:306–312, 1992.
33. Mayer, U., Nischt, R., Poschi, E., Mann, K., Gerl, M., Yamada, Y., Timpl, R. A single EGF-like motif in laminin is responsible for high affinity nidogen binding. EMBO J. 12:1879–1885, 1993.
34. Mallett, S., Fossum, S., Barclay, A.N. Characterization of the MRC OX40 antigen of activated CD4 positive lymphocytes—a molecule related to nerve growth factor receptor. EMBO J. 9:1063–1068, 1990.
35. Chalupny, N.J., Peach, R., Hollenbaugh, D., Ledbetter, J.A., Farr, A.G., Aruffo, A. T-cell activation molecule 4-1BB binds to extracellular matrix proteins. Proc. Natl. Acad. Sci. U.S.A. 89:10360–10364, 1992.