Volume 231, number 1, 1–4 FEB 05662 April 1988

Review

Structure and function of epidermal growth factor-like regions in proteins

Ettore Appella, Irene T. Weber[+] and Francesco Blasi[°]

Laboratory of Cell Biology, National Cancer Institute, Bethesda, MD 20892, [+]Center for Chemical Physics, National Bureau of Standards, Gaithersburg, MD 20899, USA and *[°]International Institute of Genetics and Biophysics, 80127 Naples, Italy and Institute for Cancer Research, Columbia University, New York, NY 10032, USA*

Received 11 January 1988

A sequence about 40 amino acid residues long which has a significant homology to epidermal growth factor (EGF) has been found in many different proteins in single or multiple copies. The homologies per se and available functional data suggest that these domains share some common functional features.

One can consider the prototype sequence as that of EGF, a 53-residue peptide, which is derived from a much longer precursor containing nine other copies of EGF-like units. EGF causes pleiotropic proliferative and developmental effects which are mediated by a specific surface receptor endowed with tyrosine kinase activity [1,2]. Certain shorter synthetic peptides compete with authentic EGF for receptor binding and mimic some of the functions of EGF, indicating that the minimal receptor binding sequence lies within residues 20–31 [3]; however, addition of the amino-flanking sequences dramatically improves its affinity for the receptor [4].

Several proteins which have significant sequence homology with EGF, or that contain one or more EGF-homologous repeats [e.g., urokinase (uPA), laminin B1 and low density lipoprotein receptor (LDL receptor)], are also involved in receptor-ligand interactions, and the sequences required for their specific interactions have been identified. uPA catalyzes the proteolytic activation of inactive plasminogen to the broad spectrum serine protease, plasmin. This reaction is regarded as crucial in regulating extracellular proteolysis, and hence in a variety of phenomena that require degradation of the extracellular matrix and of basement membrane [5]. Some normal and neoplastic cells possess a specific uPA receptor which may serve the purpose of focusing the regulatory proteolytic activity of uPA on the cell-matrix contact sites [6]. Synthetic peptide studies show that the receptor binding specificity resides within residues 18–32 of uPA, i.e., within the EGF homologous domain. As in EGF, amino-flanking sequences (residues 12–17) of uPA considerably increase the affinity for the receptor [7].

Laminin is a glycoprotein of the basement membrane that promotes the adhesion and growth of various epithelial normal and tumor cells [8]. Subunit B1 of laminin appears to be involved in binding to collagen and to a cell receptor. The sequence of laminin B1 encodes several cysteine-rich repeats with homology to EGF [9] and some of these repeats are located in domain III which contains the portion of this protein required for cell binding. Synthetic peptide studies have shown that residues 925 to 933 of laminin in one of these repeats can mediate cell attachment, migration and receptor binding [10].

Correspondence (present) address: F. Blasi, Mikrobiologisk Institut, Øster farimagsgade 2A, 1353 Copenhagen K, Denmark

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/88/$3.50 © 1988 Federation of European Biochemical Societies

The LDL receptor is a cell surface protein that binds LDL, a plasma cholesterol binding protein, and carries it into the cell [11]. The LDL receptor has three copies of a disulfide-bonded, cysteine-rich repeat of approx. 40 amino acid residues having homology to EGF. Analysis of the structure of the LDL-receptor gene in subjects with familial hypercholesterolemia has shown that deletion of exons 7 and 8 (i.e., the first two copies of the EGF-homologous repeat) strongly interferes with the binding capacity and specificity of this receptor [12].

Many other proteins contain regions of homology to EGF, and cysteine residues at similar positions (fig.1). They include tissue plasminogen activator (tPA), coagulation factors VII (with two domains), IX, X and XII (with two domains), protein C, protein S and protein Z, the *Drosophila* NOTCH sequence, the *C. elegans* LIN sequence, transforming growth factor-α and the EGF precursor, which contains other EGF-like sequences in addition to EGF. Thus, all the proteins listed above (see also fig.1) are known, or may be expected to participate in protein-protein or protein-cell interactions. They are growth factors, receptors, or receptor-like proteins or proteins of the coagulation and fibrinolytic pathway, suggesting that the EGF-like regions are responsible for the interactions with a receptor or a ligand.

In order to gain further insight into the structural characteristics of the EGF-like region, the sequence homologies in these regions of the above proteins (fig.1) must be re-evaluated. The homologies can be grouped into three different categories. In fig.1, we have numbered the cysteine residues as Cys1–Cys6, following the order in which they appear in uPA, with Cys1 the closest to the N-terminus. We note that five of the six cysteine residues, which correspond to positions Cys2–Cys6 of uPA, align very well among the different proteins, but that the Cys1 residue does not. In some proteins the extra cysteine residue is N-terminal to Cys2 (as in EGF, uPA and others), but in other cases (as in the LDL receptor) it is present between Cys3 and Cys4 (see below). We want to emphasize that, so far, the actual disulfide bonding has been determined only for EGF and TGF-α (i.e., Cys1 with Cys3, Cys2 with Cys4 and Cys5 with Cys6) [13,14].

We have grouped the sequences homologous to

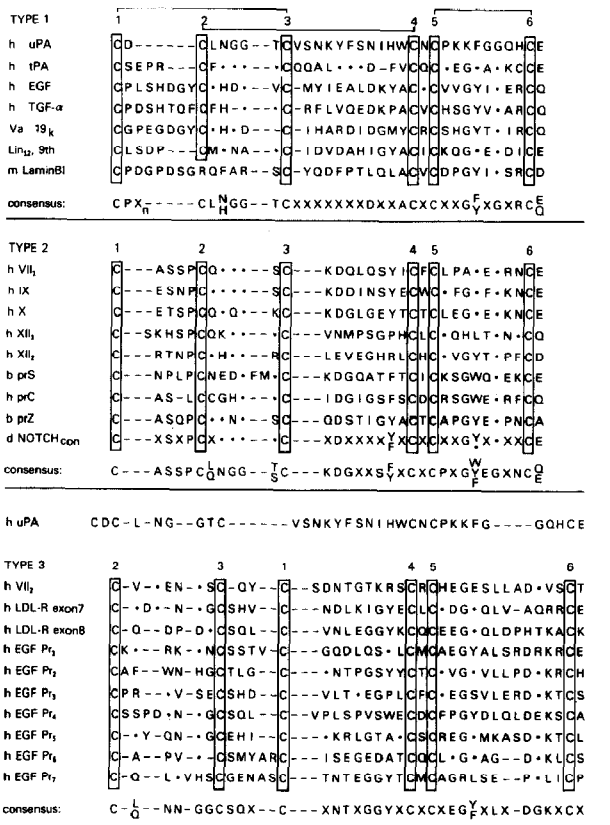

Fig.1. Sequence homology among proteins sharing an EGF-like region. Symbols follow the one-letter code. A minus sign indicates a gap introduced for sequence alignment; a dot indicates identity to the first sequence, human urokinase (h uPA). The conserved cysteines are numbered 1 to 6, and the disulfide bonds are expected to form as shown: Cys1-3, Cys2-4, Cys5-6. Sequences have been obtained as follows: h uPA [18]; human tissue plasminogen activator (h tPA) [19]; human epidermal growth factor (h-EGF) and seven repeats of the human EGF precursor (h EGF Pr_{1-7}) [20]; human transforming growth factor-α (h TGF-α) [21]; *vaccinia* virus 19 kDa protein (Va 19K) [22]; mouse laminin B1 (m $LaminB_1$) [9]; the 9th repeat of the *C. elegans* partial sequence of Lin12 (Lin_{12}, 9th) [23]; human coagulation factor VII, 1st and 2nd repeat (h VII_1 and h VII_2) [24]; human coagulation factor IX (h IX) [25]; human coagulation factor X (h X) [26]; 1st and 2nd repeat of human coagulation factor XII (h XII_1 and h XII_2) [27]; bovine protein S (b prS) [28]; human protein C (h prC) [29]; bovine protein Z (b prZ) [30]; *Drosophila melanogaster* NOTCH gene cysteine-rich repeat consensus sequence (d $NOTCH_{con}$) [31]; human LDL-receptor exons 7 and 8 (h LDL-R exon7, h LDL-R exon8) [12].

EGF as follows: in type 1 sequences (uPA, tPA, EGF, TGF-α, 19 kDa protein), the position of Cys1 is N-terminal to Cys2, with 1–7 residues between Cys1 and Cys2. The regularity in cysteine

spacing identifies three regions, A, B and C. Region A (Cys2–Cys3) is highly conserved in length (5–6 residues) and sequence. Region B (Cys3–Cys4) contains between 10 and 11 residues and is highly divergent in sequence (except in the closely related uPA and tPA). Region C (between Cys5 and Cys6) is conserved in length and has 4–5 specific residues. In the sequence of laminin B1, the Cys2 residue is missing. It is possible that a cysteine lying outside this region forms a disulfide bond with Cys4.

In type 2 sequences (coagulation factors VII_I, IX, X, $XII_{1,2,}$, protein S, protein C, protein Z, the *Drosophila* 'NOTCH' sequence), region A has the same properties as in type 1 sequences; and Cys1 is located N-terminal to Cys2 with variable separation. Also region C has features similar to those of type 1 sequences. Region B, however, is consistently three residues shorter than in type 1 sequences and has more conserved residues.

Type 3 sequences (LDL receptor, coagulation factor VII_{II}, various units of the EGF precursor) differ in several respects from the other two. Region A is less conserved in length and sequence. Region B is interrupted by Cys1. The three regions are less conserved in spacing but still contain many identifiable sequence identities with respect to each other or to type 1 and type 2. It is not clear whether a Cys1 to Cys3 disulfide bond forms in type 3 sequences, and therefore the structure may be quite different.

The comparisons in fig.1 show that consensus sequences can be generated in the three groups of proteins. Thus homology is not limited to the cysteine residues but also includes the conservation of other amino acids. Regions A and C are well conserved even in type 3 sequences (allowing for the various insertions). Region B is quite variable in type 1 sequences, but is more conserved in the other two groups.

The secondary structure of residues 1–48 of human EGF in solution has been recently determined by NMR measurements [15]. On the basis of that study, and assuming that the disulfide bonding is conserved in the EGF-like region of other proteins, a common three-dimensional structure can be predicted (fig.2). The three disulfide bonds constrain the structure to fold into four loops which correspond to the variable length region between Cys1 and Cys2, and the regions A, B and C shown in fig.2. This common structure consists of two antiparallel β-strands between Cys3 and Cys4 (loop B). The other loops have less regular secondary structure, although loop C contains two turns. This analysis suggests that several of the conserved residues are important for forming turns in the structure. In general, there is a conserved proline after Cys1 and proline or glycine just before Cys2. Loop A has a conserved pair of glycines in the middle, and loop C has two glycines, one in each of the two turns. Loop B, despite the well-defined secondary structure, does not contain conserved turn-forming residues, most likely because this loop forms the recognition site.

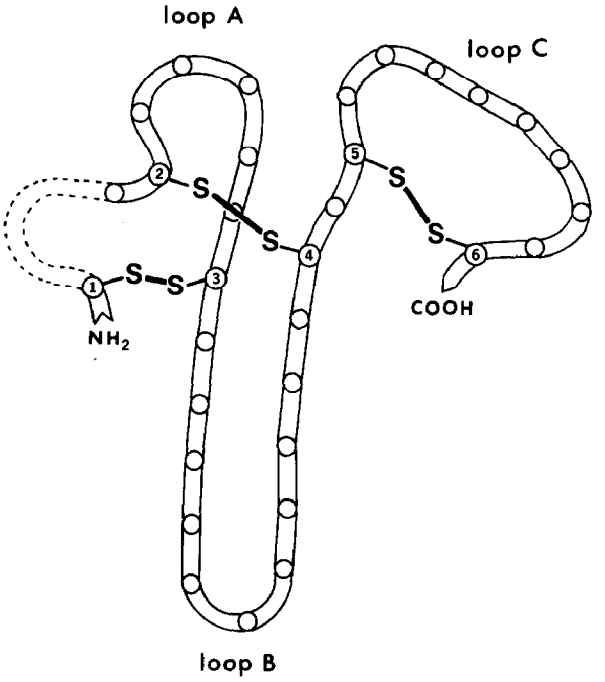

Fig.2. Predicted structure of type 1 EGF-like domains. A schematic drawing of the common structure predicted for the EGF-like domains based on the solution structure of human EGF (residues 1–48) reported by Cooke et al. [15]. The open circles represent amino acid residues. Cys1–Cys6 are numbered and the three disulfide bridges indicated. Loops A, B and C are conserved in length, while the loop between Cys1 and Cys2 (in dashed lines) is variable in length. Loop B consists of two β-strands and forms the recognition sequence. The type 2 EGF domain is predicted to have a similar structure, except that loop B is three residues shorter.

Type 1 and type 2 EGF homologous regions are predicted to fold as shown in fig.2; type 3, however, is expected to form different structures since the Cys1 and Cys3 disulfide, if present,

would be in different positions relative to type 1 and type 2 sequences. In the case of the type 1 sequences, the available data suggest that loop B is the binding site. In the case of type 2 sequences, however, since the degree of conservation is quite high in all three loops, it would be likely that additional sequences (not included in fig.1) determine or contribute to the binding site and/or specificity. Recent data support this prediction. Coagulation factor IX has been shown to have specific receptors on the surfaces of endothelial cells [16]. Short synthetic peptides derived from the sequence of the first EGF-like region (involving residues 47–54 just before the EGF-like regions of fig.1) function as competitive inhibitors of the factor IX-endothelial cell interaction [17].

The above analysis predicts: (i) that the EGF-like regions of various proteins are involved in receptor-ligand interactions; (ii) that a common structural folding is shared by the EGF-like domains, based on the sequence homologies and dictated by the position of the three disulfide bonds; (iii) that in type 1 sequences this common structure forms three loops which together provide specificity and maximum binding affinity – in these proteins, according to the data available on EGF and uPA, the recognition sequence is located in loop B, but the combination of more than one loop is probably required for maximum affinity; and (iv) that specific receptors have co-evolved with the above EGF-like structures.

For EGF, the receptor has been shown to be a membrane-spanning protein with an outer, EGF-binding domain. The EGF-like regions of the other proteins might define various classes of receptors or ligands sharing a common secondary structure, at least in the region involved in receptor-ligand interaction.

REFERENCES

[1] Carpenter, G. and Cohen, S. Ann. Rev. Biochem. 48, 193–216, 1979.
[2] Heldin, C.-H. and Westermark, H. Cell 37, 9–20.
[3] Komoriya, A., Hortsch, M., Meyers,, C., Smith, M., Kanety, H. and Schlessinger, Y. Proc. Natl. Acad. Sci. USA 81, 1351–1355, 1984.
[4] Heath, W.F. and Merrifield, B. Proc. Natl. Acad. Sci. USA 83, 6367–6371, 1986.
[5] Blasi, F., Vassalli, J.-D. and Danø, J. Cell Biol. 104, 801–804, 1987.
[6] Stoppelli, M.P., Tacchetti, C., Cubellis, M.V. Corti, A., Hearing, V.J., Cassani, G., Appella, E. and Blasi, F. Cell, 45, 675–684, 1986.
[7] Appella, E., Robinson, E.A. Ullrich, S.J., Stoppelli, M.P., Corti, A., Cassani, G. and Blasi, F. J. Biol. Chem, 262, 4437–4440, 1987.
[8] Timpl, R. Trends in Biol, Sci. 8, 207–209.
[9] Sasaki, M., Kato, S., Kohno, K., Martin, G.R. and Yamada, Y. Proc. Natl. Acad. Sci. USA 84, 935–939, 1987.
[10] Graf, J., Iwamoto, Y., Sasaki, M., Martin, G.R., Kleinman, H.K., Robey, F.A. and Yamada, Cell. 48, 990–996, 1987.
[11] Russell, D.W., Lehman, M.A., Sudhof, T.C., Yamamoto, T., Davis, C.G., Hobbs, H.H., Brown, M.S. and Goldstein, J.L. Cold Spring Harbor Symp. Quant. Biol. 51, 811–819, 1986.
[12] Yamamoto, T., Bishop, R.W., Brown, M.S., Goldstein, J.L. and Russell, D.W. Science 232, 1230–1237, 1986.
[13] Savage, C.R. Jr, Hash, J.H. and Cohen, S. J. Biol. Chem. 248, 7669–7672, 1973.
[14] Winkler, J. Biol. Chem. 261, 13838–13843, 1986.
[15] Cooke, R.M., Wilkinson, A.J., Baron, M., Pastore, A., Tappin, M.J., Campbell, I.D., Gregory, H. and Sheard, B. Nature 327, 339–341, 1987.
[16] Rimon, S., Melamed, R., Savion, N., Scot, T., Nawroth, P.P. and Stern, D.M. J. Biol. Chem., 262, 6023–6031, 1987.
[17] Nawroth, P.P., Wilner, G. and Stern, D.M. Circulation 74, 232, abstract 929.
[18] Guenzler, W.A., Steffens, G.J., Otting, F., Kim, S.A., Frankus, E. and Flohe, L. Hoppe Seylers Z. Physiol. Chem. 363, 1155–1165, 1982.
[19] Pennica, D., Holmes, W.E., Kohr, W.J., Harkins, R.N., Vehar, G.A.,. Ward, C.A., Bennett, W.F., Yelverton, E., Seeburg, P.H., Heyneker, H.L. and Goeddel, D.V. Nature 301, 214–220, 1983.
[20] Bell, G.I., Fong, N.M., Stempien, M.M., Wormsted, M.A., Caput, D., Ku, L., Urdea, M.S., Rall, S.B. and Sanchez-Pescador, R. Nucleic Ac. Res. 14, 8427–8446, 1986.
[21] Derynck, R., Roberts, A.B., Winkler, M.E., Chen, E.Y. and Goessel, D.V. Cell 38, 287–297, 1984.
[22] Venkatesan, S., Gershowitz, A. and Moss, B. J. Virol. 44, 637–646, 1982.
[23] Greenwald, I. Cell 43, 583–590, 1985.
[24] Hagen, F.S., Gray, C.L., O'Hara, P., Grant, F.J., Saari, G.C., Woodbury, R.G., Hart, C.E., Insley, M., Kisicl, W., Kurachi, K. and Davie, E.W. Proc. Natl. Acad. Sci. USA 83, 2412–2416, 1986.
[25] Yoshitake, S. Schach, B.G., Foster, D.C., Davie, E.W. and Kurachi, K. Biochemistry 24, 3736–3750, 1985.
[26] Leitus, S.P., Foster, D.C., Kurachi, K. and Davie, E.W. Biochemistry 25, 5098–5102, 1986.
[27] McMullen, B.A. and Fujikawa, K. J. Biol. Chem. 260, 5328–5341, 1985.
[28] Dahlback, B., Lundvall, Å. and Stenflo, J. Proc. Natl. Acad. Sci. USA 83, 4199–4203, 1986.
[29] Foster, D.C., Yoshitaki, S. and Davie, E.W. Proc. Natl. Acad. Sci. USA 82, 4673–4677, 1985.
[30] Højrup, P., Jensen, M.S. and Petersen, T.E. FEBS Letters 184, 333–338, 1985.
[31] Wharton, K.A., Johansen, K.M., Xu, T. and Artavanis-Tsakonas, S. Cell 43, 567–581, 1985.