Disulfide Bond Structure of Human Epidermal Growth Factor Receptor*

(Received for publication, December 8, 1997, and in revised form, February 28, 1998)

Yoshito Abe‡, Masafumi Odaka§, Fuyuhiko Inagaki, Irit Lax, Joseph Schlessinger†, and Daisuke Kohda**

From the ‡Department of Structural Biology, Biomolecular Engineering Research Institute, Furuiedai 6-chome, Suita, Osaka 565, the §Department of Molecular Physiology, The Tokyo Metropolitan Institute of Medical Science, Honhomboume 3-chome, Bunkyo-ku, Tokyo 113, Japan, and the †Department of Pharmacology, New York University Medical Center, New York, New York 10016

The extracellular domain of the human epidermal growth factor receptor (sEGFR) consists of 621 amino acid residues, including 50 cysteines. The connections of the 25 disulfide bonds in the recombinant sEGFR protein, obtained from Chinese hamster ovary cells, have been determined using N-terminal sequencing and matrix-assisted laser desorption/ionization mass spectrometry. We identified a basic repeat of eight cysteines with a 1–3, 2–4, 5–6, and 7–8 disulfide pairing pattern in the two cysteine-rich regions of sEGFR. By comparison to other cysteine-rich motifs, it was concluded that the cysteine-rich repeat of sEGFR belongs to the laminin-type EGR-like (LE) structural motif. Three-dimensional structure models of the two cysteine-rich regions have been built, based on the three-dimensional structures of the LE domains from the laminin γ1 chain and secondary structure predictions for the EGF receptor.

The epidermal growth factor receptor (EGFR)† is the cell membrane receptor for epidermal growth factor (EGF) (1). The EGFR also binds other ligands that contain amino acid sequences classified as the EGF-like motif. Among these ligands, the three-dimensional structures of EGF and transforming growth factor α (TGFα) have been determined by NMR (2). Upon binding of the ligand to the extracellular domain, the EGFR undergoes dimerization, which eventually leads to the activation of its cytoplasmic protein tyrosine kinase (1). The EGFR is also known as the ErbB-1 receptor and belongs to the type I family of receptor tyrosine kinases (1). This group also includes the ErbB-2, ErbB-3, and ErbB-4 receptors. The ligand of ErbB-2 is still unknown but that of ErbB-3 and ErbB-4 is heregulin (3). This factor is also known as neuregulin or NDF.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Present address: Biochemical Systems Laboratory, The Institute of Physical and Chemical Research (RIKEN), Hirokawa 2-1, Wako, Saitama 351-01, Japan.

** To whom correspondence should be addressed: NMR Group, Dept. of Structural Biology, Biomolecular Engineering Research Institute, Furuiedai 6-chome, Suita, Osaka 565-0874, Japan. Tel.: 06-672-8218; Fax: 06-672-8219; E-mail: kohda@beri.co.jp.

† The abbreviations used are: EGF, epidermal growth factor receptor; EG, epidermal growth factor; sEGFR, the recombinant soluble extracellular domain of EGFR; INSR, insulin receptor; LE, laminin-type EGF-like; MALDI-MS, matrix-assisted laser desorption/ionization mass spectrometry; PTH, phenylthiodydantoin; TNF, tumor necrosis factor; TNFR, tumor necrosis factor receptor; TGFα, transforming growth factor α; HPLC, high pressure liquid chromatography.

and contains an EGF-like sequence that was found to fold into an EGF-like fold by NMR (4, 5). The type II family of receptor tyrosine kinases consists of the insulin receptor (INSR), the insulin-like growth factor I receptor, and the insulin receptor-related receptor (1). Although the type II receptors have a very different quaternary structure (tetrameric αβ2) from that of the type I receptors (monomeric α), the extracellular portions of the two families, as well as the tyrosine kinase portions, share significant sequence homology, suggesting a common evolutionary origin (1, 6).

The 621 amino acid residues of the extracellular domain of the human EGFR (sEGFR) can be subdivided into four domains as follows: L1, S1, L2, and S2, where L and S stand for “large” and “small” domains, respectively (Ref. 6, see Fig. 2). L1 (residues 1–165) and L2 (residues 310–481) are homologous and are also known as domain I and domain III. Domain III (L2) is the major ligand-binding domain, since the isolated L2 domain can bind EGF and TGFα with an affinity similar to sEGFR (7, 8). L1 may have much weaker affinity for ligands, but this interaction should be important in ligand-mediated receptor dimerization (8). S1 (residues 166–309) and S2 (residues 482–621) are homologous cysteine-rich domains and are called domain II and domain IV, respectively. The number of cysteines and the spacings between the cysteines are conserved among the members of the EGFR and INSR families (6). Weaker, but significant, homologies were detected in other cysteine-rich sequences of furin (9), the TNF receptor, and laminin (10). The S1 and S2 domains of the EGFR appear to be arranged as three repeats of eight cysteines (6), but information other than sequence homology is necessary to improve the analysis of the EGFR and INSR cysteine-rich sequences.

Ligand-induced dimerization was first reported for the EGFR receptor (1) and now is widely accepted as a general mechanism for the transmission of growth stimulatory signals across the cell membrane. Although many biochemical experiments have been performed to reveal the molecular mechanism of receptor dimerization (8, 11, 12), the molecular mechanism by which monomeric ligands induce dimerization is still unknown for members of the EGFR family. Single particle averaging of electron microscopic images showed that the overall shape of the sEGFR is four-lobed and doughnut-like (12). Small angle x-ray scattering suggested that the sEGFR is a flattened sphere with long diameters of 110 Å and a short diameter of 20 Å (8). The crystallization of sEGFR in complex with EGF was published (13), but the structure has not yet been reported, despite a decade of effort by many groups. Therefore, we have started a more fundamental characterization of the EGFR, which will eventually lead to the structure determination.

In this report, we describe the connections of the 25 disulfide bonds in the extracellular domain of human EGFR. These
disulfide linkage data were essential to analyze the repeat structure of the two cysteine-rich domains of the EGFR. We concluded that the cysteine-rich repeat of EGFR is a new member of the laminin-type EGF-like (LE) motif.

**EXPERIMENTAL PROCEDURES**

**Production of the EGFR Extracellular Domain**—The extracellular domain of human EGFR (sEGFR) was produced by overexpression using Chinese hamster ovary cells, as described (7). Concentrated serum-free conditioned medium was dialyzed in 20 mM sodium phosphate buffer, pH 7.1 (buffer A), and then loaded on tandem-connected, buffer A-equilibrated columns of DEAE-Toyopearl 650S (inner diameter, 1 × 10 cm, Tosoh), CM-Toyopearl 650S (inner diameter, 1 × 10 cm, Tosoh), and Affi-Gel Blue Gel (inner diameter, 1 × 4 cm, Bio-Rad). The sEGFR was not retained by the DEAE and CM columns and was absorbed by the Affi-Gel Blue resin. The protein was eluted from the disconnected Affi-Gel Blue column with a linear gradient of 0.05 to 2 M NaCl in buffer A. The eluant containing sEGFR (~2 mg/ml) was stored at –20 °C until use.

**Assay of Free SH Group**—A 400-μl aliquot of a 6,6'-dithiodionicotinic acid solution (Wako Pure Chemical, saturated in 50 mM sodium acetate, pH 5.0, containing 8 mM guanidine HCl) was mixed with 100 μl of sample solution. The absorbance at 345 nm was measured. For calibration, 2-mercaptoethanol was used.

**Cyanogen Bromide Degradation**—The stock solution of sEGFR was desalted by gel filtration chromatography on a G-25 column (Amersham Pharmacia Biotech) equilibrated with 10% acetic acid. After lyophilization, the protein (1–5 mg) was dissolved in 100 μl of 70% formic acid and was mixed with 100 μg of cyanogen bromide. The reaction mixture was gently stirred overnight at room temperature. After the addition of 900 μl of H2O, the cyanogen bromide was removed by lyophilization.

**HPLC Separation of Peptides**—Peptides were separated by reversed-phase HPLC on a Pegasil C4 (120 Å, inner diameter 4.6 × 150 mm, Senshu Scientific Co.) column using an LC-10A system (Shimadzu) at a flow rate of 0.5 ml/min. A linear gradient was applied using solvent A containing 0.1% trifluoroacetic acid in water and solvent B containing 0.1% trifluoroacetic acid in 80% acetonitrile.

**Protease Digestion**—The purified cyanogen bromide fragments were further cleaved overnight at 37 °C in 50 mM sodium phosphate buffer, pH 6.0, by various proteases, including L-1-tosylamido-2-phenylethyl chloromethyl ketone-treated trypsin (Sigma), V8 protease (Wako), and elastase (Boehringer Mannheim). The digestion by 1-chloro-3- tosyl-amido-7-amino-2-heptanone-treated chymotrypsin (Sigma) was carried out under the same conditions, except at pH 6.5. The digestion by lysyl endopeptidase (alias Achromobacter protease I, Wako) was done overnight at 30 °C and at pH 6.0 in the presence of 2 mM urea. The digestion by trypsin and elastase (Seikagaku Corp.) was carried out for 1 h at 37 °C and at pH 6.5. In all cases, the ratio of the peptide to the protease was 10:1 (w/w).

**Treatment with Neuraminidase and Glycosidases**—The purified proteolytic fragments were treated with neuraminidase (also known as sialidase, Boehringer Mannheim), glycopeptidase A (Seikagaku Corp.), or endoglycosidase F/N-glycanase F (Boehringer Mannheim). The lyophilized glycopeptidase A powder was dissolved in 500 μl of H2O and was stored at –20 °C. The peptide solution (100 μg/10 μl in 50 mM sodium acetate, pH 5.0) was mixed with 1 μl of the neuraminidase stock solution, and the mixture was incubated for 1 h at 37 °C. For the glycopeptidase A digestion, the peptide solution (10 μg/10 μl in 50 mM sodium acetate, pH 5.0) was mixed with 10 μl of the glycopeptidase A stock solution, and the mixture was incubated overnight at 37 °C. The peptide solution (10 μg/10 μl in 50 mM sodium phosphate, pH 5.0, containing 25 mM sodium citrate, 25 mM NaCl, and 0.05% SDS) was mixed with 10 μl of the endoglycosidase F/N-glycanase F stock solution, and the mixture was incubated overnight at 37 °C.

**N-terminal Sequencing**—Peptides were sequenced on an Applied Biosystems model 492 Protein Sequencer. The Edman degradation in analyses were carried out using a Voyager Elite spectrometer (PerSeptive Biosystems). The accelerating voltage was set to 20 kV. Data were acquired in the positive linear mode of operation. Spectra were externally calibrated with angiotensin II (m/z 1046.19, Sigma) and myoglobin from horse heart (m/z 16950.70, Sigma). Protein solutions were mixed with an equal volume of matrix solution (3,5-dimethoxy-4-hydroxycinnamic acid saturated in 0.1% trifluoroacetic acid and acetonitrile, 2:1 v/v). For peptides with molecular weights <5,000, α-cyano-4-hydroxycinnamic acid saturated in 0.1% trifluoroacetic acid and acetonitrile (1:1 v/v) was used as the matrix solution. Analyses of disulfide-linked peptides were facilitated by spontaneous cleavage of the disulfide bonds during the MALDI-MS measurements (15, 16). Pseudomolecular ions corresponding to reduced forms of the peptides were observed in addition to the molecular ion at the laser fluence above the threshold.

**Protein Data Bases and Programs**—Protein sequences were obtained from release 34 of the SWISS-PROT data base (17). The sequence alignment was done by the program ClustalW (18), and final adjust-
ments were made manually. Secondary structure predictions were performed using the programs PredictProtein (19), GOR (20), nnpredict (21), and PSA (22). The module data base (23),2 release 14.0 of the Prosite data base (24), and the May, 1997, release of the SCOP data base (25) were used for collecting information on cysteine-rich motifs.

Homology modeling was done using Quanta96/Protein Design (Molecular Simulations Inc.). The sequences were aligned manually using the Sequence Viewer. The first criterion for aligning sequences was to match the conserved cysteine residues. Gaps were inserted to equalize the lengths of the inter-cysteine spacings. The dissimilarities between noncysteine residues were not considered seriously. Matched residues were selected manually, and the coordinates for the matched residues were copied from a known structure to generate a framework of a homology model. The model at this stage contained missing coordinates for atoms of some side chains and peptide segments that have no counterpart. The tools in the model backbone palette, “Regularizing Regions” and “Search Fragment Data base,” were used to find appropriate values for unknown coordinates. Then the side chain conformations were refined automatically with the tools in the model side chains palette. After connection of disulfide bonds with the CHARMMrtf option, the model was subjected to energy minimization of CHARMM in the RTF mode. The model of S22 was built first on the basis of the laminin g1III4 structure (1TLE, model 2). The models of other repeats were constructed on the basis of the S22 model structure. The relative orientations of the repeats were modeled after the manner of the laminin g1III3–5 structure (1KLO). The helix conformation between S11 and S12 was set by the tool in the model backbone palette, “Apply Conformation.”

RESULTS

The recombinant extracellular domain of the human EGF receptor was purified from Chinese hamster ovary cell culture conditioned medium. Fifty mg of protein were recovered from 3 liters of conditioned medium. N-terminal sequencing confirmed the signal cleavage site after residue Ala24 in the uncleaved protein. Therefore, the numbering starts with Leu 25 in the uncleaved protein as residue 1 in the cleaved protein (see Fig. 2). The extracellular domain of the EGFR (sEGFR) contains 50 cysteines. Free sulfhydryls were measured with 6,6-9-dithio- nicotinic acid under acidic pH conditions in the presence of 8 M guanidine HCl. The absence of an increase of absorption at 345 nm indicated the presence of less than 0.1 mol of free SH group/mol of the protein. The sEGFR molecule behaves as a monomer under nondenaturing conditions (26), indicating an absence of intermolecular disulfide bonds. Therefore all 50 cysteine residues in the sEGFR are involved in 25 intramolecular disulfide bonds.

To map the disulfide bonds, sEGFR, without reduction and deglycosylation, was cleaved by cyanogen bromide at the methionyl peptide bonds and was separated by reversed-phase HPLC (Fig. 1A). Note that all of the fractions, except for C1, gave broad and multiple peaks owing to the attached carbohydrate chains. Each fraction was analyzed by N-terminal sequencing and MALDI-MS (Table I). Intensive enzymatic deglycosylation was carried out prior to the MALDI-MS analyses to obtain the correct masses of the peptides without the carbohydrate chains. The two cysteines, Cys7 and Cys 34, were located in the fragment of peak C4. Disulfide bond formation was confirmed by the observation of di-PTH-cystine at the seventh position. 2 Available at the following on-line address: http://www.bork.embl-heidelberg.de/Modules/.
cycle of Edman degradation. The remaining cysteine residues were contained in peak C3, which was treated with glycopeptidase A for the (only partial) removal of the carbohydrate chains and then was digested by lysyl endoproteinase. The resultant peptide mixture was separated by the second reversed-phase HPLC (Fig. 1B). Analyses of 12 of the 17 peaks revealed the existence of five disulfide bonds. Each fraction contained one disulfide bond at most; hence, the identification of each disulfide linkage was straightforward. The other five peaks were processed further by various proteases to obtain fragments containing fewer disulfide bonds. Since peak C3-L15 had a very broad shape, suggesting a high content of carbohydrate, the fraction was treated with neuraminidase and then with glycosidase F/N-glycanase prior to tryptic digestion. During the analyses after the third fragmentation, nine disulfide bonds were positively identified, since each fragment contained only one disulfide bond. If a fragment contained more than one disulfide bond, the assignment was slightly complicated. For example, peak C3-L7-P1 contained four cysteines in three peptides. The three peptides must be linked by two disulfide bonds to one another, since the mass peak corresponding to the sum of the three peptides was observed in the MALDI-MS spectrum. Thus, there remained two possible disulfide linkages (Scheme 1).

Di-PTH-cystine was observed at the fifth and seventh cycles of the Edman degradation, clearly demonstrating that the combination on the left was correct. If the combination on the right had been correct, then di-PTH-cystine would have been observed at the second cycle, instead of the fifth. Another example was found in peak C3-L10-TE3. This peak comprised three peptide chains, but only two chains were found during the N-terminal sequencing. The mass value obtained by the prompt fragmentation during MALDI-MS showed that the third peptide was Gln<sup>893</sup>Ser<sup>896</sup>, but it must have been dehydrated (Table I). This anomaly is probably due to the formation of pyroglutamic acid at the N-terminal glutamine residue of the third peptide. Attempts to remove the pyroglutamic acid by pyroglutamate aminopeptidase (Boehringer Mannheim), however, were unsuccessful. The observation of di-PTH-cystine at the sixth cycle of the Edman degradation indicated the disulfide bond connection as in Table I.

Peak C3-L12-E2 was subjected to the fourth fragmentation by chymotrypsin and subsequently by prolyl endopeptidase to generate only one peak, C3-L12-E2-CP1, containing two disulfide bonds. Peaks C3-L15-T1 and C3-L15-T2 were derived from the second cysteine-rich domain and therefore contained many cysteine residues. Repeated intensive digestions using less specific proteases, such as elastase and prolyl endopeptidase, were necessary to obtain fragments with one or two disulfide bonds. After the fourth and fifth fragmentations, seven disulfide bonds were identified eventually (Table I).

In summary, 338 residues were confirmed by N-terminal sequencing, accounting for 54% of the sEGFR residues. The recovered peptide fragments identified by the combination of N-terminal sequencing and MALDI-MS contained a total of 524 residues, accounting for 84% of the sEGFR. The rest of the residues were ascribed to peptides that were difficult to isolate, due to their small size (<5 residues). Some long peptides were not analyzed simply because they did not have disulfide bonds. All but one of the disulfide bonds were positively identified by the Edman degradation. The last disulfide bond, between...
Cys212 and Cys224, was assigned by elimination because of the unsuccessful recovery of the fragment containing the disulfide bond (see the line labeled as “missing” in Table I). It is likely that the peptide fragment became very small after the intensive digestion with elastase, which is less specific.

**DISCUSSION**

The disulfide bonds in the recombinant human EGF receptor have been determined by a combination of N-terminal sequencing and mass spectroscopy. Peptide cleavage, deglycosylation, and purification were carried out at acidic pH values and under nonreducing conditions to prevent disulfide scrambling. No indication of disulfide rearrangement was observed during the analysis. The combination of all 25 intramolecular disulfide linkages in the EGF receptor is summarized in Fig. 2.

Domains L1 and L2 (domains I and III, respectively) have homologous amino acid sequences, and each contains four conserved cysteines. The disulfide pairing pattern in the L domains was found to be 1–2,3–4. Domains S1 and S2 (domains II and IV, respectively) have also homologous and contain many cysteines, i.e., 22 in S1 and 20 in S2. The 11 disulfide bonds in S1 are arranged into three units of a 1–3,2–4 pattern and five units of a 1–2 pattern, and the 10 disulfides in S2 are arranged into three units of a 1–3,2–4 pattern and four units of a 1–2 pattern. The cysteine-rich sequences of EGFR and INSR are classified into the furin Cys-rich motif group (23). Some members, but not all, of the mammalian subtilisin-related proteases that cleave precursor proteins at dibasic sequences, such as furin and PACE4, contain many repeats of eight cysteines (10, 27). A relationship between the EGFR Cys-rich repeat and the laminin-type EGF-like repeat was also pointed out, and the disulfide bonding pattern was predicted based on the working hypothesis that the eight cysteine repeat appeared to be a chimera of a TNF receptor repeat and an EGF-like repeat (10).

Surprisingly, this prediction is correct despite the rough assumption, except for two disulfide bonds in the segment of residues 287–309. They suggested a 1–4,2–3 pairing pattern, i.e., Cys287-Cys309 and Cys302-Cys305, but in fact, Cys287-Cys302 and Cys305-Cys309 are the correct combinations.
structures were available: EGF, TGF sequences for which three-dimensional from the alignment using the EGF-like. Those for the EGF-like motif were made from the alignment using the EGF-like sequences for which three-dimensional structures were available: EGF, TGF, herregulin-α, EGF-like modules of fibrillin, coagulation factors X and IX, E-selectin, P-selectin, thrombomodulin, and prostaglandin H2 synthetase-1. B, pattern of the EGF-like module defined in the Prosite database: EGF_1 (Prosite PS00022) and EGF_2 (Prosite PS01186). C, patterns of the Cys-rich repeats of EGFR obtained from the multiple alignment given in Fig. 4.

Fig. 5. Cysteine spacings and conserved non-cysteine residues in cysteine-rich motifs. A, comparison of the spacings between the conserved cysteines. The ranges of the spacings for the EGFR/INSR/furin, laminin type EGF-like, and TNFR Cys-rich motifs were derived from the multiple alignments (10). Those for the EGF-like motif were made from the alignment using the EGF-like sequences for which three-dimensional structures were available: EGF, TGF, herregulin-α, EGF-like modules of fibrillin, coagulation factors X and IX, E-selectin, P-selectin, thrombomodulin, and prostaglandin H2 synthetase-1. B, pattern of the EGF-like module defined in the Prosite database: EGF_1 (Prosite PS00022) and EGF_2 (Prosite PS01186). C, patterns of the Cys-rich repeats of EGFR obtained from the multiple alignment given in Fig. 4.

Fig. 6. Schematic representation of the folds of selected cysteine-rich motifs. A, laminin-type EGF-like fold (Protein Data Base codes 1TLE and 1KLO); B, EGF-like fold (1EPG and 3EGF); C, TNF receptor Cys-rich fold (1TNF); D, EGFR Cys-rich fold (model, this study). Cysteine residues are drawn as open circles, and non-cysteine residues contained in fixed-length spacings between the conserved cysteines are drawn as smaller closed circles. Hydrogen bonds are drawn by dotted lines.

Disulfide Structure of EGF Receptor

families on the basis of the multiple alignment of the cysteine-rich repeats (Fig. 4).

Many cysteine-rich motifs are found in various extracellular proteins. There are presently more than 30, and new cysteine-rich motifs are continually being identified (23). Such conserved sequence motifs can fold into a globular shape autonomously and often correspond to single exons (28). For more than 10 motifs, the disulfide linkage pattern and the three-dimensional structure have already been determined experimentally (23, 24). The motif of the EGFR Cys-rich repeat is most similar to the laminin-type EGF-like (LE) motif; both are repeats of eight cysteine residues, with a 1–3, 2–4, 5–6, 7–8 pairing pattern. The epidermal growth factor-like (EGF-like) motif is also a structurally related repeat of six cysteines, with a 1–3, 2–4, 5–6 pattern. This motif is seen in all ligands of the EGF receptor family. Although the EGF-like motif and the LE motif share a similar cysteine spacing pattern, and even similar structural features, as shown below, they should be treated as different motifs. The tumor necrosis factor receptor (TNF) also consists of repeats of six cysteines. The generally accepted disulfide pattern of the TNFR Cys-rich motif is 1–2, 3–5, 4–6, but this pattern is equivalent to those of the EGF-like and LE motifs after a circular permutation.

Fig. 5A shows a comparison of the lengths of the spacings between the conserved cysteines. Some spacings with fixed short lengths are unique to each Cys-rich motif, but other spacings show large variations in length. In the Prosite database, the EGF-like domain signature is defined as two consen-

sus patterns, EGF_1 and EGF_2 (Fig. 5B). If the C5-C6 spacing is short, a Gly/Pro-aromatic sequence is expected three residues before C6. If the spacing is equal to or longer than eight residues, then a Gly/Pro-aromatic sequence is expected three res-
A typical EGF-like sequence has eight residues between C5 and C6, so that the sequence shows a merged pattern like C<sub>X</sub>XX<sub>G</sub>fX<sub>G</sub>XXC, where “f” denotes an aromatic amino acid. Many LE repeats have an EGF<sub>1</sub> and/or EGF<sub>2</sub> features, and thus they are classified as laminin-type “EGF-like.” The LE motif, however, contains eight instead of six cysteine residues. In the case of the Cys-rich repeats of the EGFR family, we can extract a pattern from each Cys-rich repeat (Fig. 5) by reference to the multiple alignment (Fig. 4). The patterns in S21 and S22 have eight residues between C5 and C6, and the characteristic Gly residue, indicating that the two patterns resemble the EGF<sub>1</sub> pattern. Another pattern in S14 has 11 residues between C5 and C6 and an Asn-aromatic sequence. Since asparagine, as well as glycine, is frequently seen at the corners of β-turns, this Asn-aromatic sequence is considered to be equivalent to the Gly-aromatic sequence in the EGF<sub>2</sub> pattern. The last pattern found in S13 deviates from the typical EGF-like patterns but is regarded as an analog of the EGF<sub>1</sub> and EGF<sub>2</sub> patterns. The main difference of the EGFR Cys-rich motif from the EGF-like pattern lies in the C4-C5 spacing as follows: two residues in the EGFR Cys-rich repeats, but one residue in the EGF-like motif. Since the same length variation of the C4-C5 spacing is seen in the case of the LE motif and the TNFR Cys-rich motif (Fig. 5A), we can expand the definition of the EGF-like patterns from C4-X(1)-C5 to C4-X(1,2)-C5. Therefore, we consider the EGFR Cys-rich repeat (and also the INSR and Furin Cys-rich repeats) to be a new member of the LE motif.

The x-ray and NMR structures of the LE modules derived from the mouse laminin γ1 chain were reported recently (29, 30). The NMR structure (Protein Data Base entry, 1TLE) was the single LE module that contains the high affinity binding site for nidogen. The backbone fold of the LE module is shown schematically in Fig. 6A. The LE module can be visualized as a triple loop structure held by four disulfide bonds. The N-terminal segment containing the first three disulfide bonds folds into the same topology as the EGF-like module (Fig. 6B). A typical EGF-like fold contains two anti-parallel β-sheets, whereas the LE module has a low content of secondary structure elements (29, 30). Despite the difference in the secondary structure content, they are classified into the same superfamily named “EGF/laminin” in the SCOP data base (25), although they are separated at the next family level. The Cys-rich repeat of the TNFR also shows the same backbone fold (Fig. 6C) but clearly has a different protein architecture as follows: the positions of the two cysteine residues at C3 and C4 on the β-sheet are shifted by four residues. By contrast, the corresponding cysteine residues are adjacent on the β-sheet in the LE and EGF-

![Image](https://example.com/image.png)
like folds. We propose that the fold of the EGFR Cys-rich repeat is identical to the LE module, as shown in Fig. 6D.

The x-ray LE structure (Protein Data Base entry, 1KLO) of the mouse laminin-1 chain has the three consecutive LE modules, with the central LE corresponding to the NMR structure. The two adjacent LE modules are attached to each other by a hook-like association of the C3-C4 loop with the C7-C8 loop of the preceding LE module, so that the overall structure has a rod-like shape (30). We constructed molecular models of the S1 and S2 domains of EGFR by homology modeling on the basis of the NMR and x-ray LE structures and secondary structure predictions for the human EGFR receptor. In usual homology modeling, secondary structure elements are aligned first, and then short loops connecting the secondary structure elements are added. In contrast, however, useful anchoring points in the LE structure are only the position of cysteine residues. The conformation of long loops consisting of residues that have no counterpart was simply chosen by visual inspection among hits of the fragment data base search performed with the program Quanta96/Protein Design. Thus, the models presented below are regarded as guidance of the global structure, and the details of conformation are not meaningful.

The modeling of the S2 domain was straightforward, since the S2 domain is a simple truncated version of the x-ray laminin structure (Fig. 7B). For example, the sequence alignment and residue matching used in the model building of S22 are shown (Fig. T, inset). In contrast, special attention was paid during the modeling of the S1 domain. One point is the deletion of the N-terminal half of the fourth repeat, S14 (Fig. 3). We assume that the relative orientation of the C7-C8 loop of S13 against the C5-C6 loop of S14 is similar to that of the C5-C6 loop and the C7-C8 loop found within a single LE fold (Fig. 6A), resulting in a sharp kink in the S1 model (Fig. 7A). The other point is concerned with the conformation of the peptide segment between S11 and S12. This segment has a conserved length of seven residues in the EGF family (Fig. 4A). Secondary structure prediction was carried out using various programs on the internet. These programs found little secondary structure in the cysteine-rich domains, except for the segment between S11 and S12. The programs GOR and nnpredict predicted a short helix and a short extended strand in this segment, and the programs PredictProtein and PSA predicted a helix with three to four turns in the same region. We prefer the single helix to the helix + strand structure, because a nonpolar amino acid cluster characteristic of an a-helix is found at positions i - 3 (Cys185), i (Leu186), i + 3 (Ile189), and i + 4 (Ile190) (Fig. 4A). Polar amino acid residues at other positions suggest that this putative helix is amphipathic and lies between a hydrophobic core and the solvent. The putative helix between the S11 and S12 domains could be fitted smoothly in the S1 model structure (Fig. 7A). In the case of the Drosophila and C. elegans EGF receptors, a shorter helix was predicted (Fig. 4A). The corresponding segment in the INSR family is three residues long, and no secondary structure was suggested by the secondary structure prediction models.

The model of the S2 Cys-rich domain of the EGFR has a slightly bent rod-like shape and may function as a rigid mechanical spacer that keeps the major ligand-binding domain 2 away from the cell membrane (Fig. 7B). The predicted model has a total length of about 70 Å and a diameter of about 15 Å. By contrast, many deletions occurred in the S1 Cys-rich domain of EGFR during evolution, so the S1 domain has a much more complex organization of repeats than the S2 domain (Fig. 3). The model of the S1 Cys-rich domain has a rod-like shape, with a similar size to the S2 domain, but it kinks at the junction of S13 and S14 and contains a putative amphipathic helix between S11 and S12 (Fig. 7A). We consider that the S1 domain is not a simple mechanical spacer but that it must have more active functions. For further discussion, we need know how S1 and S2 interact with each other or with L1 and L2.

In conclusion, we have determined the disulfide bond connections in the human EGFR receptor. We propose that the two cysteine-rich domains of the EGFR are composed of repeats of the laminin-type EGF-like (LE) motif. The determination of the disulfide bonds is the first step for the structural characterization of the extracellular domain of the EGF receptor family.