54. M. A. Lane, A. Sainten, G. M. Cooper, *Cell* **28**, 873 (1982).
55. D. E. Comings, *Proc. Natl. Acad. Sci. U.S.A.* **70**, 3324 (1973).
56. A. G. Knudson, in *Cancer: Achievements, Challenges and Prospects for the 1980's*, J. H. Burchenal and H. F. Oettgen, Eds. (Grune & Stratton, New York, 1981), pp. 381–396.
57. W. F. Benedict, A. L. Murphree, A. Banerjee, C. A. Spina, M. C. Sparkes, R. S. Sparkes, *Science* **219**, 973 (1983).
58. H. P. Klinger, *Cytogenet. Cell Genet.* **27**, 254 (1980); A. B. Sabin, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 7129 (1981); E. J. Stanbridge *et al.*, *Science* **215**, 252 (1982).
59. J. J. Mulvihill *et al.*, in *Genetics of Human Cancer* (Raven New York, 1977), p. 137.
60. J. J. Yunis, *Hum. Genet.* **56**, 293 (1981).
61. C. J. Der, T. G. Krontiris, G. M. Cooper, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 3637 (1982).
62. It is possible that activation of oncogenes detected by transfection assays represent a noninitiating step in oncogenesis, since NIH 3T3 mouse cells, used in the transfection assay to identify some cellular oncogenes, represent preneoplastic rather than normal cells [E. P. Reddy, R. K. Reynolds, E. Santos, M. Barbacid, *Nature (London)* **300**, 149 (1982); J. W. Littlefield, *Science* **218**, 214 (1982)]. Expression of the malignant phenotype following integration of an oncogene in a normal cell has not been demonstrated.
63. J. D. Rowley, *Nature (London)* **243**, 290 (1973).
64. C. D. Bloomfield, L. C. Peterson, J. J. Yunis, R. D. Brunning, *Br. J. Haematol.* **36**, 347 (1977).
65. J. D. Rowley, *Ann. Genet.* **16**, 109 (1973).
66. A. Hagemeijer, K. Hahlen, W. Sizoo, J. Abels, *Cancer Genet. Cytogenet.* **5**, 95 (1982).
67. H. van den Berghe *et al*, *Blood* **48**, 624 (1976).
68. G. H. Borgström *et al.*, *Cancer Genet. Cytogenet.* **2**, 115 (1980).
69. A. A. Sandberg, *Hum. Pathol.* **12**, 531 (1981).
70. H. van den Berghe *et al.*, *Hum. Genet.* **46**, 173 (1979); M. Nagasaka *et al.*, *Blood* **61**, 1174 (1983); J. L. Parkin *et al.*, *ibid.* **60**, 1321 (1982).
71. J. Mark, R. Dahlenfors, C. Ekedahl, G. Stenman, *Cancer Genet. Cytogenet.* **2**, 231 (1980).
72. Study supported in part by NIH grants CA31024 and CA33314 and by grant 6-286 from the National Science Foundation.

RESEARCH ARTICLE

Structure of a Mouse Submaxillary Messenger RNA Encoding Epidermal Growth Factor and Seven Related Proteins

James Scott, Mickey Urdea, Margarita Quiroga
Ray Sanchez-Pescador, Noel Fong, Mark Selby
William J. Rutter, Graeme I. Bell

Abstract. *The structure of the messenger RNA (mRNA) encoding the precursor to mouse submaxillary epidermal growth factor (EGF) was determined from the sequence of a set of overlapping complementary DNA's (cDNA). The mRNA is unexpectedly large, about 4750 nucleotide bases, and predicts the sequence of preproEGF, a protein of 1217 amino acids (133,000 molecular weight). The EGF moiety (53 amino acids) is flanked by polypeptide segments of 976 and 188 amino acids at its amino and carboyxl termini, respectively. The amino terminal segment of the precursor contains seven peptides with sequences that are similar but not identical to EGF.*

Epidermal growth factor (EGF) is a 53 amino acid protein that has been isolated from the submaxillary gland of the male mouse and from human urine (*1*). It stimulates the proliferation and differentiation of cells of ectodermal and mesodermal origin. In addition, EGF, which is presumably identical to the hormone urogastrone, is a potent inhibitor of HCl release from the intestinal mucosa. As EGF exerts a number of effects on prenatal and neonatal tissue growth including accelerated maturation of the lung, precocious eye-opening, and incisor eruption and is found in elevated levels in milk, it may play a role in early development. Moreover, since EGF receptors are present in various adult tissues, EGF is presumably involved in the control of growth and function of cells throughout life.

Interestingly, EGF stimulates phosphorylation of its own receptor by a receptor-associated tyrosine-specific protein kinase which may be related to those encoded by the transforming genes of some retroviruses (*2*). Thus, the control of cell proliferation by EGF and retroviruses may share common features.

EGF is synthesized in the tubular cells of the submaxillary gland of the mouse, in the acinar cells of the human submaxillary gland, and in the human duodenal glands (*3*). Although the primary translation product of EGF messenger RNA (mRNA) has not been identified, EGF is probably generated by proteolytic processing of a larger molecule, as forms with molecular weights of about 9,000, and 28,000 and 30,000 have been reported in the mouse submaxillary gland and human urine, respectively (*4*).

We report here the nucleotide sequence of the mRNA encoding mouse submaxillary gland preproEGF. The mRNA which is at least 4750 bases encodes an EGF precursor of 1217 amino acids. The sequence of the precursor contains EGF and seven peptides that possess structural similarity to EGF.

EGF-specific clones were isolated from a male mouse submaxillary complementary DNA (cDNA) library (*5*) by hybridization with ^{32}P-labeled synthetic oligodeoxynucleotide probes made as four pools of 64-fold degenerate 20-base oligonucleotides according to the nucleotide sequence predicted from the amino acid sequence of EGF(17–23) (*6*). Eleven of 5000 colonies hybridized with the probes of pool 4; one colony, pmegf10, contained a plasmid with an insert of about 1700 base pairs (bp). This insert contained a continuous opening reading frame which included the sequence of mouse EGF. Hybridization of ^{32}P-labeled pmegf10 (*7*) to male and female mouse submaxillary RNA indicated that the mRNA encoding EGF is at least ten times more abundant in the male gland, as expected (*1*), and is approximately 4800 bases. Since the insert in pmegf10 was not a complete copy of the mRNA, overlapping clones were identified by screening the original 5000 and 7500 additional colonies with terminal restriction fragments prepared from the insert in pmegf10.

This strategy was repeated with other restriction fragments to identify all colonies that contained a portion of the mRNA

James Scott, Mark Selby, and William J. Rutter are in the Department of Biochemistry and Biophysics, University of California, San Francisco 94143. Mickey Urdea, Margarita Quiroga, Ray Sanchez-Pescador, Noel Fong, James Scott, and Graeme I. Bell (to whom requests for reprints should be sent) are on the staff of Chiron Corporation, 4560 Horton Street, Emeryville, California 94608. The present address for James Scott is Molecular Medicine Group, MRC, Clinical Research Centre, Harrow, Middlesex HA1 3UJ, United Kingdom.

encoding preproEGF. The composite sequence of the overlapping cDNA fragments from pmegf10b, 10, 1, and 44 (Fig. 1) did not include about 1000 bases from the 5′ end of the mRNA. Therefore another library was synthesized with the use of an oligonucleotide primer, 3′-CCGCTTCCTTCGGTGCGAAT-5′ (*8*), complementary to nucleotides 1032 to 1051 (Fig. 3). The sequence of the mRNA was deduced from the sequences of both sets of overlapping cDNA clones (Figs. 1 and 3). As none of the 3′ cDNA clones contained a polyadenylate [poly(A)] tract, there could be additional nucleotides at the 3′ end of this sequence. However, there is a polyadenylation signal (AAUAAA) 18 bases from the end of the sequence (nucleotide 4750) as well as another at nucleotides 4359 to 4364. The 5′-untranslated region may be nearly complete because the size of the cloned segment in pmegf39, 1051 bp, is about the size (±5 bp) of the elongated primer determined from a sequencing gel. The RNA contains a single large open reading frame of 3396 bases beginning at the first Met (*8*) codon, nucleotides 354 to 356, and encodes a protein of 1217 amino acids (molecular weight of 133,000) which includes EGF (amino acids 977 to 1029). There are termination codons in all frames upstream of the assigned initiating Met. The 5′- and 3′-untranslated regions of the mRNA are at least 353 and 746 bases, respectively. The sequence of the insert in pmegf35 differs from that in pmegf39 at one position in the 5′-untranslated region and at three in the coding region, two of which change the amino acid sequence, Val^{10} to Phe^{10} and Asp^{173} to Asn^{173}. These differences probably reflect sequence polymorphism since restriction mapping

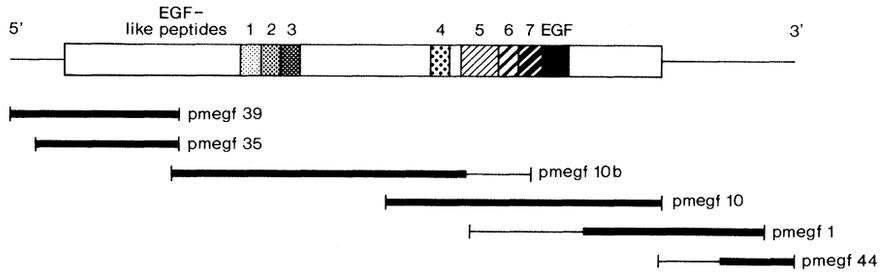


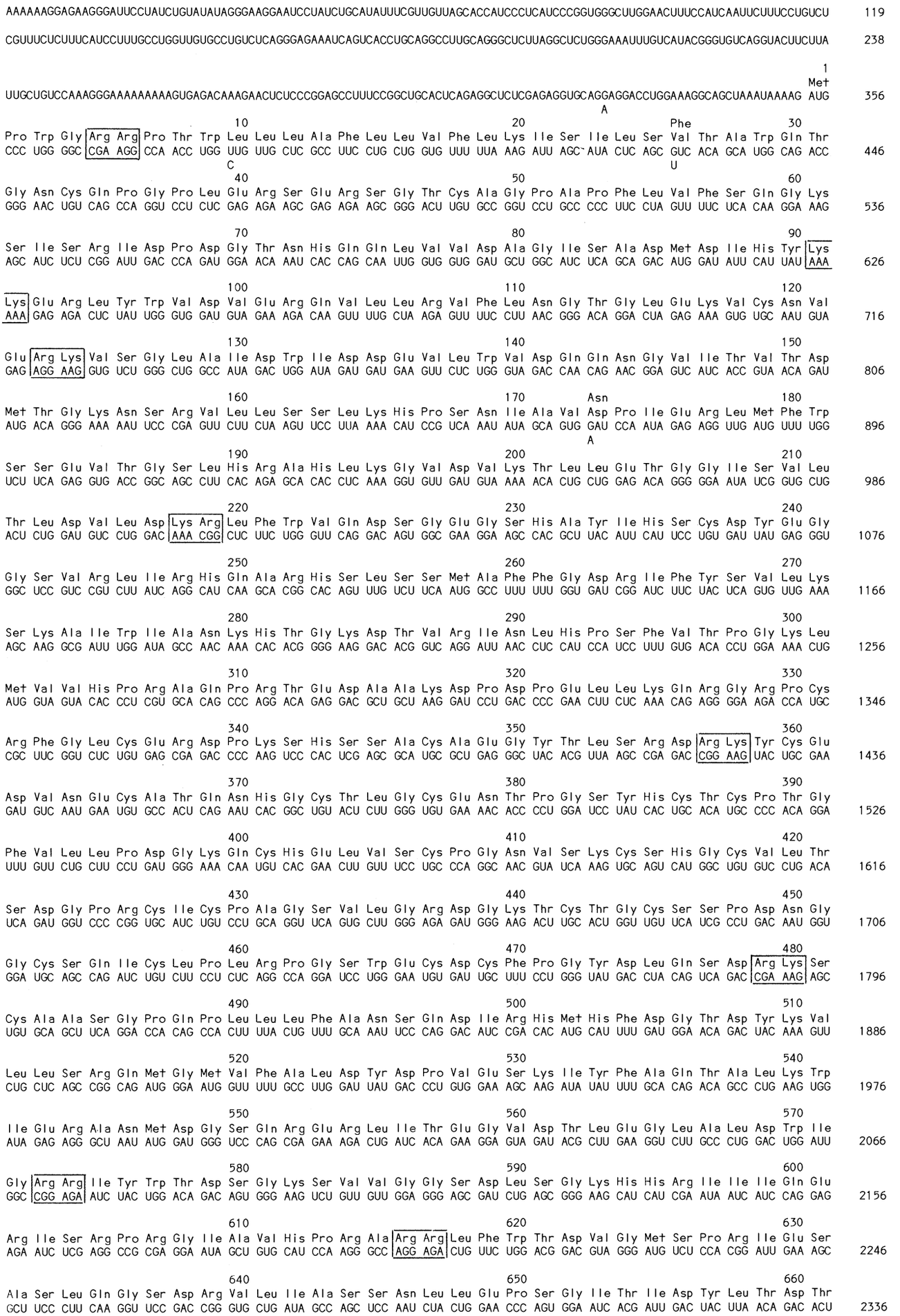
Fig. 1. Schematic representation of mouse submaxillary EGF mRNA and protein. The box indicates the protein-coding portion of the mRNA. The positions of the seven EGF-like peptides and EGF are indicated. The thin line corresponds to the untranslated regions of the mRNA. The structure of the set of overlapping cDNA clones from which the sequence was derived is indicated, and the thick lines represent the region of each cDNA which was sequenced.

```
1 (357-399)      Arg Lys Tyr [Cys] Glu Asp Val Asn Glu [Cys] Ala Thr Gln Asn His [Gly] [Cys] Thr
2 (400-440)            Gln [Cys] His Glu Leu Val Ser [Cys] Pro Gly Asn Val Ser Lys [Cys] Ser
3 (441-480)                  Thr [Cys] Thr Gly [Cys] Ser Ser Pro Asp Asn Gly Gly [Cys] Ser Gln
4 (745-784)                  Lys Pro Gly Ala Asp Pro [Cys] Leu Tyr Arg Asn Gly [Gly] [Cys] Glu
5 (803-885)  Met Val Ser Gly Met Asn Tyr Glu Asp Asp [Cys] Gly Pro Gly Gly [Cys] [Gly] Ser His
6 (886-925)                                                  Ser Asp [Cys] Pro Ser
7 (926-976)                                             Gly Ala His Asn [Cys] Ala Glu Asn
EGF (977-1029) Asn Ser Tyr Pro Gly [Cys] Pro Ser Ser Tyr Asp Gly Tyr [Cys] Leu Asn [Gly] Gly Val

Leu Gly [Cys] Glu Asn [Thr] Pro [Gly] Ser [Tyr] His [Cys] Thr [Cys] Pro Thr [Gly] Phe Val Leu
His Gly [Cys] Val Leu [Thr] Ser Asp Gly Pro Arg [Cys] Ile [Cys] Pro Ala [Gly] Ser Val Leu
Ile [Cys] Leu Pro Leu Arg Pro [Gly] Ser Trp Glu [Cys] Asp [Cys] Phe Pro [Gly] Tyr Asp Leu
His Ile [Cys] Gln Glu Ser Leu [Gly] Thr Ala Arg [Cys] Leu [Cys] Arg Glu [Gly] Phe Val Lys
Ala Arg [Cys] Val Ser Asp Gly Glu Thr Ala Glu [Cys] Gln [Cys] Leu Lys [Gly] Phe Ala Arg
Ser Arg [Cys] Ile Asn [Thr] Glu [Gly] Gly [Tyr] Val [Cys] Arg [Cys] Ser Glu [Gly] Tyr Glu Gly
Ala Ala [Cys] Thr Asn [Thr] Glu [Gly] Gly [Tyr] Asn [Cys] Thr [Cys] Ala Gly Arg Pro Ser Ser
[Cys] Met His Ile Glu Ser Leu Asp Ser [Tyr] Thr [Cys] Asn [Cys] Val Ile [Gly] Tyr Ser Gly

Leu Pro [Asp] Gly [Lys]
Gly Arg [Asp] Gly [Lys]
Gln Ser [Asp] Arg [Lys]
Ala Trp [Asp] Gly [Lys]
Asp Gly Asn Leu [Cys] Ser [Asp] Ile Asp Glu [Cys] Val Leu Ala Arg
Asp Gly Ile Ser [Cys] Phe [Asp] Ile Asp Glu [Cys] Gln Arg
Pro Gly Arg Ser [Cys] Pro [Asp] Ser Thr Ala Pro Ser Leu Leu Gly Glu Asp Gly His His Leu Asp Arg
Asp Arg [Cys] Gln Thr Arg [Asp] Leu Arg Trp Trp Glu Leu Arg
```

Fig. 2. Comparison of EGF-like peptides 1 to 7 and EGF. The proteins were aligned at the common Cys-X-Cys sequence. Identical amino acids shared by at least four members of the family and Cys residues are boxed. The boundaries of the proteins are indicated in parentheses. Only the sequence of EGF-like peptide 5 from amino acids 832 to 885 is indicated. EGF-like peptide 1 is the top line in each group.



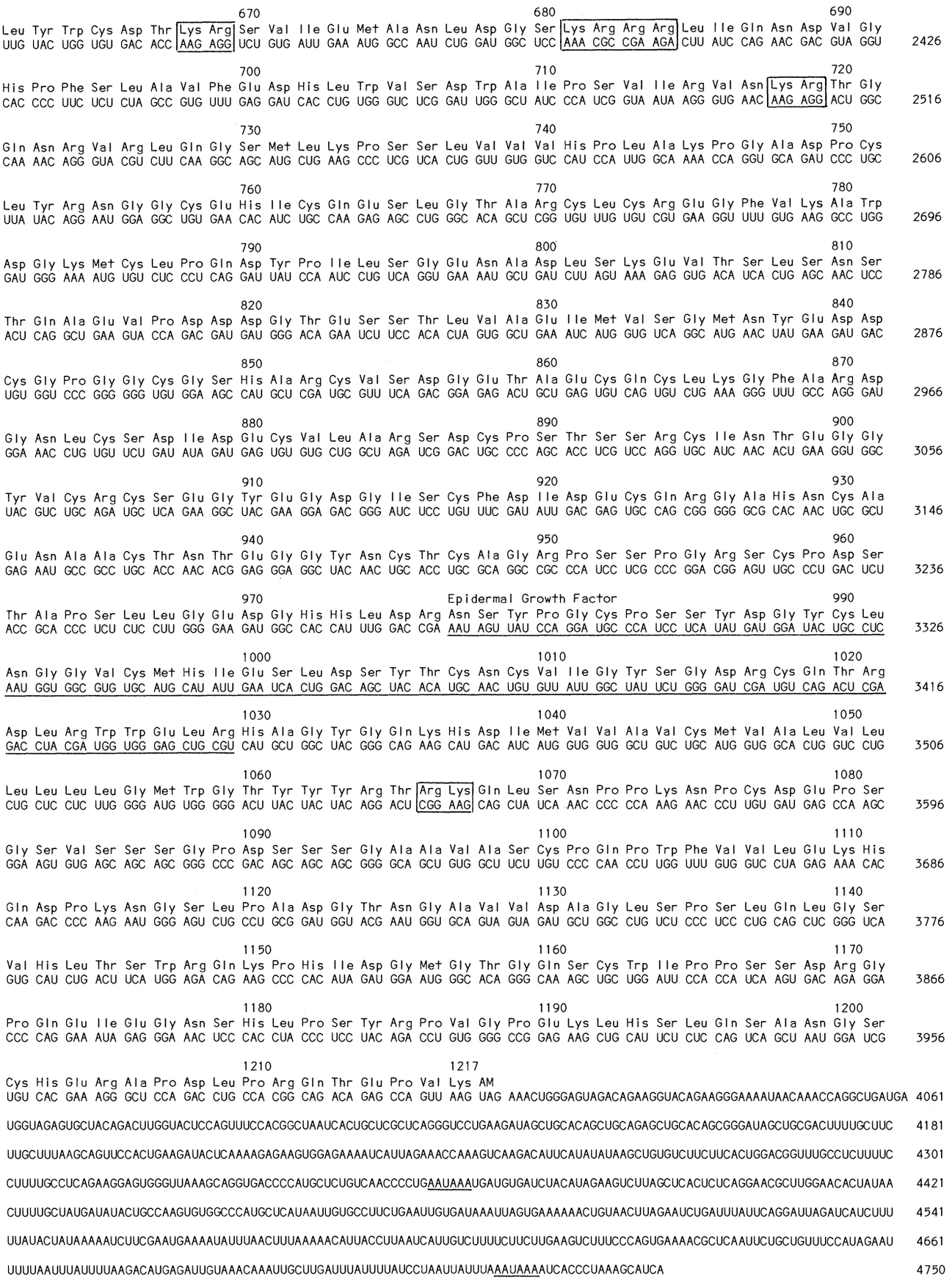


Fig. 3. The sequence of mouse submaxillary EGF precursor mRNA and protein. The predicted amino acid sequence of preproEGF is numbere by designating the first methionine as amino acid 1. The region corresponding to EGF is labeled and underlined. The positions of the seven EGI like peptides (ELP) are: ELP-1, residues 357 to 399; ELP-2, residues 400 to 440; ELP-3, residues 441 to 480; ELP-4, residues 745 to 784; ELP-5 residues 803 to 885; ELP-6, residues 886 to 925; and ELP-7, residues 926 to 976. Basic di- and tetrapeptides are boxed. The number of th nucleotide at the end of each line is indicated. The two AAUAAA sequences in the 3′-untranslated region are underlined. The DNA sequence wa determined (*14*) on both strands and across all restriction sites used to initiate sequence determination.

of eight independently isolated clones indicated that four have Asp^{173} and four have Asn^{173}.

In common with other secreted proteins, the EGF precursor probably has an amino terminal signal peptide of 15 to 25 amino acids (*9*). The amino-terminal region of the precursor contains a hydrophobic section, residues 7 to 19, which is characteristic of signal peptides. EGF, amino acids 977 to 1029, is flanked by polypeptide segments of 976 and 188 amino acids, and its sequence is identical to that determined from the protein. Thus, its release from the precursor requires proteolytic processing at both ends of the molecule. In that EGF can be isolated in association with an arginine-specific peptidase, the EGF-binding protein, and has a carboxyl-terminal Arg, it has been suggested that this esteropeptidase activity might be involved in processing of the precursor (*1, 4, 10*). This concept is supported by the sequence since there is an Arg adjacent to the amino-terminal Asn of EGF. The precursor also contains 11 pairs of basic amino acids (exclusive of one in the signal peptide region), that are often sites of proteolytic processing in other systems (*9*). It is unknown whether this activity is present in the tubular cells of the submaxillary gland. Since the processing pathway is not obvious from the sequence of proEGF, it is difficult to predict the positions of the high molecular weight forms of EGF that have been described (*4*). However, EGF(9000) could correspond to EGF with a carboxyl-terminal extension produced by cleavage at Arg 1064 or 1066.

ProEGF is unexpectedly large, and it seems unlikely that it is processed to yield a single biologically active entity of 53 amino acids. However, comparison of its sequence with those in the National Biomedical Research Foundation data bank revealed homology to only mouse and human EGF. Nor is the sequence of the EGF-binding protein (*11*) present in the preproEGF sequence. Inspection of the sequence of proEGF revealed an apparently nonrandom distribution of Cys residues. In particular, there were several occurrences of the sequence Cys-X-Cys, one within EGF and seven in the region of the precursor amino terminal to EGF. Alignment of the regions containing this sequence (Fig. 2) suggests that there may be seven cryptic peptides of different sizes within the precursor which are structurally similar, but not identical, to EGF. Besides a similarity in the position and number of Cys residues, other amino acids are identical or represent conservative replacements. Moreover, the homology is increased by insertion of gaps in the sequences. The boundaries of each EGF-like peptide can be defined by a basic amino acid and thus they could be released by a trypsin-like activity. However, the arginine-specific esteropeptidase which releases EGF is probably not sufficient because several of the putative cleavage sites are Lys residues. The sizes of the peptides vary from 39 to 83 amino acid residues and five are about 40 residues. Six of these EGF-like sequences are tandemly arranged in two groups of three members each. EGF is at the carboxyl terminus of the second group (Figs. 1 to 3). The other solitary EGF-like protein is located between the two tandem arrays and separated by a spacer of about 20 amino acids from the group containing EGF. Thus, proEGF may be a protein which is processed to yield a number of different peptides. Since the aggregate size of EGF and the EGF-like peptides, 390 amino acids, accounts for only 34 percent of the precursor, additional active peptides could also be formed from proEGF.

Protein factors have a central role in regulating growth and differentiation. The submaxillary gland of the male mouse contains unusually high concentrations of a number of growth factors, including nerve growth factor and EGF, which has facilitated their isolation and characterization (*12*). However, these growth factors probably have other physiologically significant and as yet undetermined sites of synthesis as well. Since growth factors, at least nerve growth factor (*5*) and EGF, are derived from much larger proteins, tissue or temporal-specific processing of the precursor could generate a family of proteins with different biological properties and novel effects on development. The presence of several EGF-like peptides in proEGF suggests that each may have a distinct biological role. Whether they act through their own receptor is not known. However, the possibility exists that there may be a family of receptors which bind EGF and these related peptides with different affinities.

The availability of cDNA probes will facilitate the unambiguous determination of the sites of synthesis of EGF. Since the sequence and organization of the EGF precursor is now known, polypeptides can be synthesized and used to produce antiserums so that the processing of the precursor can be critically examined. It will then be possible to determine whether the EGF-like peptides identified in proEGF are present in the submaxillary gland and in other sites of synthesis of EGF. The antiserums can also be used to detect other proteins which can be generated from the precursor. Synthesis of the EGF-like peptides will facilitate an analysis of their physiological function. In particular, it will be interesting to compare their properties with the mitogenic and gastric-acid inhibitory activities of EGF, as well as with the growth-promoting activities of the transforming growth factors (*13*), a poorly defined group of peptides, which bind to the EGF receptor.

References and Notes

1. C. R. Savage, T. Inagami, S. Cohen, *J. Biol. Chem.* **247**, 7612 (1971); H. Gregory, *Nature (London)* **257**, 325 (1975); G. Carpenter and S. Cohen, *Annu. Rev. Biochem.* **48**, 193 (1979); G. Carpenter, *Handb. Exp. Pharmacol.* **57**, 89 (1981); D. Gospodarowicz, *Annu. Rev. Physiol.* **43**, 251 (1981).
2. M. Chinkers and S. Cohen, *Nature (London)* **290**, 516 (1981); J. E. Kudlow, J. E. Buss, G. N. Gill, *ibid.*, p. 519.
3. R. W. Turkington, J. L. Males, S. Cohen, *Cancer Res.* **31**, 252 (1971); E. Gresik and T. Barka, *J. Histochem. Cytochem.* **25**, 1027 (1977); S. VanNoorden, P. Heitz, M. Kasper, A. G. E. Pearse, *Histochemistry* **52**, 329 (1977); P. U. Heitz, M. Kasper, S. VanNoorden, J. M. Polak, H. Gregory, A. G. E. Pearse, *Gut* **19**, 408 (1978).
4. P. Frey, R. Forand, T. Maciag, E. M. Shooter, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 6294 (1979); Y. Hirata and D. N. Orth, *J. Clin. Endocrinol. Metab.* **48**, 673 (1979).
5. J. Scott, M. Selby, M. Urdea, M. Quiroga, G. I. Bell, W. J. Rutter, *Nature (London)* **302**, 538 (1983).
6. Oligonucleotides were synthesized by solid-phase phosphoramidate methodology [S. L. Beaucage and M. H. Caruthers, *Tetrahedron Lett.* **22**, 1859 (1981)]. The sequences in each pool were: pool 1, 3′-CCNCCNCANACA-TACGTGTA-5′ (N indicates each base was present); pool 2, 3′-CCNCCNCANACATACG-TATA-5′; pool 3, 3′-CCNCCNCANACG-TACGTGTA-5′; pool 4, 3′-CCNCCNCAN-ACGTACGTATA-5′.
7. P. S. Thomas, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 5201 (1980).
8. Abbreviations for the bases are: A, adenine; C, cytosine; G, guanine; U, uracil. Abbreviations for the amino acid residues are: Ala, alanine; Arg, arginine; Asp, aspartic acid; Asn, asparagine; Cys, cysteine; Glu, glutamic acid; Gln, glutamine; His, histidine; Ile, isoleucine; Leu, leucine; Lys, lysine; Met, methionine; Phe, phenylalanine; Pro, proline; Ser, serine; Thr, threonine; Trp, tryptophan; Tyr, tyrosine; Val, valine.
9. D. F. Steiner, P. S. Quinn, S. J. Chan, J. Marsh, H. S. Tager, *Ann. N.Y. Acad. Sci.* **343**, 1 (1980); D. D. Sabatini, G. Kreibich, T. Morimoto, M. Adesnik, *J. Cell Biol.* **92**, 1 (1982).
10. J. M. Taylor, W. M. Mitchell, S. Cohen, *J. Biol. Chem.* **249**, 3198 (1974); A. C. Server, A. Sutter, E. M. Shooter, *ibid.* **251**, 1188 (1976).
11. K. A. Thomas, N. C. Baglan, R. A. Bradshaw, *ibid.* **256**, 9156 (1981).
12. T. Barka, *J. Histochem. Cytochem.* **28**, 836 (1980).
13. R. Baserga, Ed., *Tissue Growth Factors* (Springer-Verlag, New York, 1981); A. B. Roberts *et al.*, *Nature (London)* **295**, 417 (1982); L. M. Matrisian, M. Pathak, B. E. Magun, *Biochem. Biophys. Res. Commun.* **107**, 761 (1982); M. B. Sporn, A. B. Roberts, J. H. Shull, J. M. Smith, J. M. Ward, J. Sodek, *Science* **219**, 1329 (1983).
14. A. M. Maxam and W. Gilbert, *Methods Enzymol.* **65**, 499 (1980); F. Sanger, A. R. Coulson, B. G. Barrell, A. J. H. Smith, B. A. Roe, *J. Mol. Biol.* **143**, 161 (1980).

15. We thank L. B. Rall, P. Valenzuela, and M. Appling for their assistance. Supported in part by NIH grant 21344 (W.J.R.) and a grant from the European Molecular Biology Organization (J.S.).

17 May 1983