- 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
- 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.
- W. F. Benedict, A. L. Murphree, A. Banerjee, C. A. Spina, M. C. Sparkes, R. S. Sparkes, *Science* 219, 973 (1983).
- 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., 58.
- Science 215, 252 (1982). J. J. Mulvihill et al., in Genetics of Human Cancer (Raven New York, 1977), p. 137. 59.
- J. Yunis, Hum. Genet. 56, 293 (1981).
 C. J. Der, T. G. Krontiris, G. M. Cooper, Proc. Natl. Acad. Sci. U.S.A. 79, 3637 (1982).
 It is possible that activation of oncogenes detected by transfection assays represent a non-initiating step in oncogenesis, since NIH 3T3 mouse cells, used in the transfection assay to identify some cellular oncogenes represent identify some cellular oncogenes, represent represent of the second of the malignant phenotype following integration of an oncogene in a normal cell has not been demonstrated.
- G. J. D. Rowley, *Nature (London)* 243, 290 (1973).
 C. D. Bloomfield, L. C. Peterson, J. J. Yunis, R.

- D. Brunning, Br. J. Haematol. 36, 347 (1977).
 55. 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).
- G. H. Borgström et al., Cancer Genet. Cyto-genet. 2, 115 (1980).
 A. A. Sandberg, Hum. Pathol. 12, 531 (1981).
 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) 69. (1982)
- J. Mark, R. Dahlenfors, C. Ekedahl, G. Sten-71 man, *Cancer Genet.* Cytogenet. **2**, 231 (1980). Study supported in part by NIH grants CA31024 and CA33314 and by grant 6-286 from the Na-72. tional Science Foundation.

RESEARCH ARTICLE

cessing 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 ³²P-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 ³²P-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

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

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 mesocontrol 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

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.

dermal 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

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 pro-

James Scott, Mark Selby, and William J. Rutter are in the Department of Biochemistry and Biophys-ics, 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



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)	A	rg L	ys 1	ſyr	Cys	Glu	Asp	Val	Asn	Glu	Cys	Ala	Thr	Gln	Asn	His	Gly	Cys	Thr
2	(400-440)			G	Əln	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 V	al So	er G	Эlу	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 Giy Cys Giu Asn Thr Pro Giy Ser Tyr His Cys Thr Cys Pro Thr Giy Phe Val Leu His Giy Cys Val Leu Thr Ser Asp Giy Pro Arg Cys IIe Cys Pro Ala Giy Ser Val Leu IIe Cys Leu Pro Leu Arg Pro Giy Ser Trp Giu Cys Asp Cys Phe Pro Giy Tyr Asp Leu His IIe Cys Gin Giu Ser Leu Giy Thr Ala Arg Cys Leu Cys Arg Giu Giy Phe Val Lys Ala Arg Cys Val Ser Asp Giy Giu Thr Ala Giu Cys Gin Cys Leu Lys Giy Phe Ala Arg Ser Arg Cys IIe Asn Thr Giu Giy Giy Tyr Val Cys Arg Cys Ser Giu Giy Tyr Giu Giy Ala Ala Cys Thr Asn Thr Giu Giy Giy Tyr Asn Cys Thr Cys Ala Giy Arg Pro Ser Ser Cys Met His IIe Giu Ser Leu Asp Ser Tyr Thr Cys Asn Cys Val IIe Giy Tyr Ser Giy

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. EGFlike peptide 1 is the top line in each group.

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 $(\pm 5 \text{ bp})$ of the elongated

primer determined from a sequencing

gel. The RNA contains a single large

open reading frame of 3396 bases begin-

ning 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 ac-

ids 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 posi-

tion in the 5'-untranslated region and at

three in the coding region, two of which

change the amino acid sequence, Val¹⁰

to Phe¹⁰ and Asp¹⁷³ to Asn¹⁷³. These

differences probably reflect sequence

polymorphism since restriction mapping

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 Cyś Gln Thr Arg Asp Leu Arg Trp Trp Glu Leu Arg

AAAAAAGGAGAAGGGAUUCCUAUCUGUAUAUAUAGGGAAGGAA													SUCU	119															
CGUUU	GUUUCUCUUUCAUCCUUUGCCUGGUUGUGCCUGUCUCAGGGAGAAAUCAGUCACCUGCAGGCCUUGCAGGGCUCUUAGGCUCUGGGAAAUUUGUCAUACGGGUGUCAGGUACUUCUUA															CUUA	238												
UUGCUG	1 Met IGCUGUCCAAAGGGAAAAAAAGUGAGACAAAGAACUCUCCCGGAGCCUUUCCGGCUGCACUCAGAGGCUCUCGAGAGGGUGCAGGAGGACCUGGAAAGGCAGCUAAAUAAA															1 Me† AUG	356												
Pro Tr CCC UC	-p GI GG GG	y Arg C CGA	Arg AGG	Pro CCA	Thr ACC	Trp UGG	10 Leu UUG	Leu UUG	Leu CUC	Ala GCC	Phe UUC	Leu CUG	Leu CUG	Val GUG	Phe UUU	Leu UUA	20 Lys AAG	lle AUU	Ser AGC	A Ile AUA	Leu CUC	Ser AGC	Phe Val GUC	Thr ACA	Ala GCA	Trp UGG	30 Gln CAG	Thr ACC	446
GIY As	sn Cy AC UG	s GIn	Pro	Gly	Pro	Leu	C 40 Glu GAG	Arg AGA	Ser	Glu	Arg Aga	Ser	Gly	Thr	Cys	Ala	50 Giy	Pro	Ala	Pro	Phe	Leu	U Val GUU	Phe	Ser	GIn	60 Giy	Lys	536
Ser I	le Se	r Arg	lle	Asp	Pro	Asp	70 Gly	Thr	Asn	His	GIn	GIn	Leu	Val	Val	Asp	80 Ala	Gly	lle	Ser	Ala	Asp	Met	Asp	lle	His	90 Tyr	Lys	
Lys G	u Ar	g Leu	Tyr	Trp	Val	Asp	100 Val	Glu	AAU	GIn	Val	Leu	Leu	Arg	Val	GAU Phe	110 Leu	Asn	GIy	Thr	GUA	Leu	Glu	Lys	Val	Cys	120 Asn	Val	626
<u>AAA</u> GA	G AG	A CUC	UAU Ser	UGG	GUG Leu	GAU	GUA 130	GAA	AGA	CAA	GUU	UUG	GLU	AGA Val	GUU	UUC	CUU 140 Val	AAC	GGG	ACA Gln	GGA	GIV	GAG Val	AAA	GUG	Val	AAU 150 Thr	GUA	716
GAG AC	G AA	e ene	UCU	GGG	CUG	GCC	AUA 160	GAC	UGG	AUA	GAU	GAU	GAA	GUU	CUC	UGG	GUA 170	GAC	CAA	CAG Asn	AAC	GGA	GUC	AUC	ACC	GUA	ACA 180	GAU	806
AUG AC	CA GG	G AAA	ASN	UCC	CGA	GUU	CUU	CUA	AGU	UCC	UUA	AAA	CAU	CCG	UCA	ASN	AUA 200	GCA	GUG	Asp GAU A	CCA	AUA	GAG	AGG	UUG	AUG	210	UGG	896
Ser Se UCU UC	er GI CA GA	u Val G GUG	Thr ACC	G I y GGC	Ser AGC	Leu CUU	His CAC 220	Arg AGA	Ala GCA	His CAC	Leu CUC	Lys AAA	G I y GGU	Val GUU	Asp GAU	Val GUA	Lys AAA 230	Thr ACA	Leu CUG	Leu CUG	Glu GAG	Thr ACA	G I y GGG	GIy GGA	lle AUA	Ser UCG	Val GUG 240	Leu CUG	986
Thr Le ACU CU	eu As JG GA	p Val U GUC	Leu CUG	Asp GAC	Lys AAA	Arg CGG	Leu CUC	Phe UUC	Trp UGG	Val GUU	Gln CAG	Asp GAC	Ser AGU	G I y GGC	Glu GAA	G I y GGA	Ser AGC	His CAC	Ala GCU	Tyr UAC	lle AUU	His CAU	Ser UCC	Cys UGU	Asp GAU	Tyr UAU	Glu GAG	G I y GGU	1076
GIY Se GGC UC	er Va CC GU	I Arg C CGU	Leu CUU	lle AUC	Arg AGG	His CAU	GIn CAA	Ala GCA	Arg CGG	His CAC	Ser AGU	Léu UUG	Ser UCU	Ser UCA	Me† AUG	Ala GCC	260 Phe UUU	Phe UUU	G I y GGU	Asp GAU	Arg CGG	lle AUC	Phe UUC	Tyr UAC	Ser UCA	Val GUG	270 Leu UUG	Lys AAA	1166
Ser Ly AGC A/	/s Al AG GC	a lle G AUU	Trp UGG	lle AUA	Ala GCC	Asn AAC	280 Lys AAA	His CAC	Thr ACG	G I y GGG	Lys AAG	Asp GAC	Thr ACG	Val GUC	Arg AGG	lle AUU	290 Asn AAC	Leu CUC	His CAU	Pro CCA	Ser UCC	Phe UUU	Val GUG	Thr ACA	Pro CCU	G I y GGA	300 Lys AAA	Leu CUG	1256
Met Va AUG Gl	al Va JA GU	I His NA CAC	Pro CCU	Arg CGU	Ala GCA	Gln CAG	310 Pro CCC	Arg AGG	Thr ACA	Glu GAG	Asp GAC	Ala GCU	Ala GCU	Lys AAG	Asp GAU	Pro CCU	320 Asp GAC	Pro CCC	G I u GAA	Leu CUU	Ĺeu CUC	Lys AAA	Gln CAG	Arg AGG	G I y GGA	Arg AGA	330 Pro CCA	Cys UGC	1346
Arg Pi CGC Ul	ne GI JC GG	y Leu W CUC	Cys UGU	G I u GAG	Arg CGA	Asp GAC	340 Pro CCC	Lys AAG	Ser UCC	His CAC	Ser UCG	Ser AGC	Ala GCA	Cys UGC	Ala GCU	G I u GAG	350 Gly GGC	Tyr UAC	Thr ACG	Leu UUA	Ser AGC	Arg CGA	Asp GAC	Arg CGG	Lys AAG	Tyr UAC	360 Cys UGC	Glu GAA	1436
Asp Va GAU GL	al As JC AA	n Glu U GAA	Cys UGU	Ala GCC	Thr ACU	Gln CAG	370 Asn AAU	His CAC	G I y GGC	Cys UGU	Thr ACU	Leu CUU	G I y GGG	Cys UGU	G I u GAA	Asn AAC	380 Thr ACC	Pro CCU	G I y GGA	Ser UCC	Tyr UAU	His CAC	Cys UGC	Thr ACA	Cys UGC	Pro CCC	390 Thr ACA	G I y GGA	1526
Phe Va	al Le	u Leu	Pro	Asp	Gly	Lys	400 GIn	Cys	His	Glu	Leu	Val	Ser	Cys	Pro	Gly	410 Asn	Val	Ser	Lys	Cys	Ser	His	Gly	Cys	Val	420 Leu	Thr	1616
Ser As	sp GI	y Pro	Arg	Cys	lle	Cys	430 Pro	Ala	Gly	Ser	Val	Leu	Gly	Arg	Asp	Gly	440 Lys	Thr	Cys	Thr	Gly	Cys	Ser	Ser	Pro	Asp	450 Asn	Gly	1010
UCA GA	∖U GG ∕s Se	er Gin	CGG	UGC Cvs	AUC	UGU Pro	CCU 460 Leu	GCA Ara	GGU Pro	UCA	GUG Ser	CUU	GGG G I u	AGA Cys	GAU	GGG Cys	AAG 470 Phe	ACU Pro	UGC	ACU Tyr	GGU Asp	UGU Leu	UCA	UCG Ser	CCU Asp	GAC	AAU 480 Lys	GGU Ser	1706
GGÁ UG	GC AG	IC CAG	AUC	UĠU	CUU	CCU	CUC 490	AGĞ	CCA	GGÁ	UCC	UGĠ	GAA	UGU	GAÙ	UĠC	500	CCU	GGĠ	UÁU	GAĊ	CUA	CAG	UCA	GAC	CGA	510	AGC	1796
UGU GO	CA GC	U UCA	GGA	CCA	CAG	CCA	CUU 520	UUA	CUG	UUU	GCA	AAU	UCC	CAG	GAC	AUC	CGA	CAC	AUG	CAU	UUU	GAU	GGA	ACA	GAC	UAC	AAA 540	GUU	1886
Leu Le CUG CI	eu Se JC AG	er Arg GC CGG	Gln CAG	Me† AUG	G I y GGA	Me† AUG	Val GUU 550	Phe UUU	Ala GCC	Leu UUG	Asp GAU	Tyr UAU	Asp GAC	Pro CCU	Val GUG	Glu GAA	Ser AGC 560	Lys AAG	lle AUA	Tyr UAU	Phe UUU	Ala GCA	Gln CAG	Thr ACA	Ala GCC	Leu CUG	Lys AAG 570	Trp UGG	1976
lle G AUA G/	lu Ar AG AG	g Ala G GCU	Asn AAU	Me† AUG	Asp GAU	G I y GGG	Ser UCC 580	Gln CAG	Arg CGA	Glu GAA	Arg AGA	Leu CUG	lle AUC	Thr ACA	Glu GAA	GIy GGA	Val GUA 590	Asp GAU	Thr ACG	Leu CUU	Glu GAA	G I y GGU	Leu CUU	Ala GCC	Leu CUG	Asp GAC	Trp UGG 600	lle AUU	2066
GIY A GGC C	ng Ar GG AG	g Ile A AUC	Tyr UAC	Trp UGG	Thr ACA	Asp GAC	Ser AGU	GIy GGG	Lys AAG	Ser UCU	Val GUU	Val GUU	G I y GGA	Gly GGG	Ser AGC	Asp GAU	Leu CUG	Ser AGC	GIy GGG	Lys AAG	His CAU	His CAU	Arg CGA	lle AUA	lle AUC	lle AUC	GIn CAG	Glu GAG	2156
Arg I AGA AI	le Se JC UC	er Arg G AGG	Pro CCG	Arg CGA	GIy GGA	lle AUA	ь10 Аlа GCU	Val GUG	His CAU	Pro CCA	Arg AGG	Ala GCC	Arg AGG	Arg AGA	Leu CUG	Phe UUC	620 Trp UGG	Thr ACG	Asp GAC	Val GUA	G I y GGG	Me† AUG	Ser UCU	Pro CCA	Arg CGG	lle AUU	G I U G A A	Ser AGC	2246
Ala So GCU UG	er Le CC CU	eu Gin IU CAA	G I y GGU	Ser UCC	Asp GAC	Arg CGG	640 Val GUG	Leu CUG	lle AUA	Ala GCC	Ser AGC	Ser UCC	Asn AAU	Leu CUA	Leu CUG	Glu GAA	650 Pro CCC	Ser AGU	G I y GGA	lle AUC	Thr ACG	lle AUU	Asp GAC	Tyr UAC	Leu UUA	Thr ACA	660 Asp GAC	Thr ACU	2336

Leu UUG	Tyr UAC	Trp UGG	Cys UGU	Asp GAC	Thr ACC	Lys AAG	Arg AGG	670 Ser UCU	Val GUG	lle AUU	Glu GAA	Me† AUG	Ala GCC	Asn AAU	Leu CUG	Asp GAU	G I y GGC	680 Ser UCC	Lys AAA	Arg CGC	Arg CGA	Arg AGA	Leu CUU	lle AUC	Gln CAG	Asn AAC	Asp GAC	690 Val GUA	G I y GGU	2426
His CAC	Pro CCC	Phe UUC	Ser UCU	Leu CUA	Ala GCC	Val GUG	Phe UUU	700 Glu GAG	Asp GAU	His CAC	Leu CUG	Trp UGG	Val GUC	Ser UCG	Asp GAU	Trp UGG	Ala GCU	710 Ile AUC	Pro CCA	Ser UCG	Val GUA	lle AUA	Arg AGG	Val GUG	Asn AAC	Lys AAG	Arg AGG	720 Thr ACU	G I y GGC	2516
G I n CAA	Asn AAC	Arg AGG	Val GUA	Arg CGU	Leu CUU	Gln CAA	G I y GGC	730 Ser AGC	Met AUG	Leu CUG	Lys AAG	Pro CCC	Ser UCG	Ser UCA	Leu CUG	Val GUU	Val GUG	740 Val GUC	His CAU	Pro CCA	Leu UUG	Ala GCA	Lys AAA	Pro CCA	G I y GGU	Ala GCA	Asp GAU	750 Pro CCC	Cys UGC	2606
Leu UUA	Tyr UAC	Arg AGG	Asn AAU	G I y GGA	G I y GGC	Cys UGU	G I u GA A	760 His CAC	lle AUC	Cys UGC	GIn CAA	Glu GAG	Ser AGC	Leu CUG	G I y GGC	Thr ACA	Ala GCU	770 Arg CGG	Cys UGU	Leu UUG	Cys UGU	Arg CGU	G I u GA A	G I y GGU	Phe UUU	Val GUG	Lys AAG	780 Ala GCC	Trp UGG	2696
Asp GAU	G I y GGG	Lys AAA	Me† AUG	Cys UGU	Leu CUC	Pro CCU	Gln CAG	790 Asp GAU	Tyr UAU	Pro CCA	lle AUC	Leu CUG	Ser UCA	G I y GGU	Glu GAA	Asn AAU	Ala GCU	800 Asp GAU	Leu CUU	Ser AGU	Lys AAA	G I u GAG	Val GUG	Thr ACA	Ser UCA	Leu CUG	Ser AGC	810 Asn AAC	Ser UCC	2786
Thr ACU	Gln CAG	Ala GCU	G I u GA A	Val GUA	Pro CCA	Asp GAC	Asp GAU	820 Asp GAU	G I y GGG	Thr ACA	Glu GAA	Ser UCU	Ser UCC	Thr ACA	Leu CUA	Val GUG	Ala GCU	830 Glu GAA	lle AUC	Met AUG	Val GUG	Ser UCA	G I y GGC	Me† AUG	Asn AAC	Tyr UAU	G I u GA A	840 Asp GAU	Asp GAC	2876
Cys UGU	G I y GGU	Pro CCC	G I y GGG	G I y GGG	Cys UGU	G I y GGA	Ser AGC	850 His CAU	Ala GCU	Arg CGA	Cys UGC	Val GUU	Ser UCA	Asp GAC	G I y GGA	G I u GAG	Thr ACU	860 Ala GCU	G I u GAG	Cys UGU	Gln CAG	Cys UGU	Leu CUG	Lys AAA	Gly GGG	Phe UUU	Ala GCC	870 Arg AGG	Asp GAU	2966
G I y GGA	Asn AAC	Leu CUG	Cys UGU	Ser UCU	Asp GAU	lle AUA	Asp GAU	880 Glu GAG	Cys UGU	Val GUG	Leu CUG	Ala GCU	Arg AGA	Ser UCG	Asp GAC	Cys UGC	Pro CCC	890 Ser AGC	Thr ACC	Ser UCG	Ser UCC	Arg AGG	Cys UGC	lle AUC	Asn AAC	Thr ACU	Glu GAA	900 G I y GGU	G I y GGC	3056
Tyr UAC	Val GUC	Cys UGC	Arg AGA	Cys UGC	Ser UCA	G I u GAA	G I y GGC	910 Tyr UAC	G I u GAA	G I y GGA	Asp GAC	G I y GGG	lle AUC	Ser UCC	Cys UGU	Phe UUC	Asp GAU	920 e AUU	Asp GAC	G I u GAG	Cys UGC	Gln CAG	Arg CGG	G I y GGG	Ala GCG	His CAC	Asn AAC	930 Cys UGC	Ala GCU	3146
G I u GAG	Asn AAU	Ala GCC	Ala GCC	Cys UGC	Thr ACC	Asn AAC	Thr ACG	940 Glu GAG	G I y GGA	G I y GGC	Tyr UAC	Asn AAC	Cys UGC	Thr ACC	Cys UGC	Ala GCA	G I y GGC	950 Arg CGC	Pro CCA	Ser UCC	Ser UCG	Pro CCC	G I y GGA	Arg CGG	Ser AGU	Cys UGC	Pro CCU	960 Asp GAC	Ser UCU	3236
Thr ACC	Ala GCA	Pro CCC	Ser UCU	Leu CUC	Leu CUU	G I y GGG	Glu GAA	970 Asp GAU	G I y GGC	His CAC	His CAU	Leu UUG	Asp GAC	Arg CGA	Epic Asn <u>AAU</u>	derm Ser AGU	al Gi Tyr UAU	rowtł Pro CCA	h Fa Gly GGA	ctor Cys UGC	Pro CCA	Ser UCC	Ser UCA	Tyr UAU	Asp GAU	G I y GGA	Tyr UAC	990 Cys UGC	Leu CUC	3326
Asn AAU	G I y GGU	G I y GGC	Val GUG	Cys UGC	Me† AUG	His CAU	lle AUU	100 G I u GAA	0 Ser UCA	Leu CUG	Asp GAC	Ser AGC	Tyr UAC	Thr ACA	Cys UGC	Asn AAC	Cys UGU	1010 Val GUU) Ile AUU	G I y GGC	Tyr UAU	Ser UCU	G I y GGG	Asp GAU	Arg CGA	Cys UGU	Gln CAG	102 Thr ACU	0 Arg CGA	3416
Asp <u>GAC</u>	Leu CUA	Arg CGA	Trp UGG	Trp UGG	G I u GAG	Leu CUG	Arg CGU	103 His CAU	0 Ala GCU	G I y GGC	Tyr UAC	G I y GGG	Gln CAG	Lys AAG	His CAU	Asp GAC	lle AUC	1040 Me† AUG) Val GUG	Val GUG	Ala GCU	Val GUC	Cys UGC	Me† AUG	Val GUG	Ala GCA	Leu CUG	105 Val GUC	D Leu CUG	3506
Leu CUG	Leu CUC	Leu CUC	Leu UUG	G I y GGG	Me† AUG	Trp UGG	G I y GGG	106 Thr ACU	0 Tyr UAC	Tyr UAC	Tyr UAC	Arg AGG	Thr ACU	Arg CGG	Lys AAG	Gln CAG	Leu CUA	1070 Ser UCA	0 Asn "AAC	Pro CCC	Pro CCA	Lys AAG	Asn AAC	Pro CCU	Cys UGU	Asp GAU	G I u GAG	108 Pro CCA	0 Ser AGC	3596
G I y GGA	Ser AGU	Val GUG	Ser AGC	Ser AGC	Ser AGC	G I y GGG	Pro CCC	109 Asp GAC	0 Ser AGC	Ser AGC	Ser AGC	G I y GGG	Ala GCA	Ala GCU	Val GUG	Ala GCU	Ser UCU	110 Cys UGU	0 Pro CCC	Gln CAA	Pro CCU	Trp UGG	Phe UUU	Val GUG	Val GUC	Leu CUA	G I u GAG	111 Lys AAA	0 His CAC	3686
Gln CAA	Asp GAC	Pro CCC	Lys AAG	Asn AAU	G I y GGG	Ser AGU	Leu CUG	112 Pro CCU	0 Ala GCG	Asp GAU	G I y GGU	Thr ACG	Asn AAU	G I y GGU	Ala GCA	Val GUA	Val GUA	113 Asp GAU	0 Ala GCU	G I y GGC	Leu CUG	Ser UCU	Pro CCC	Ser UCC	Leu CUG	Gln CAG	Leu CUC	114 Gly GGG	0 Ser UCA	3776
Val GUG	His CAU	Leu CUG	Thr ACU	Ser UCA	Trp UGG	o Arg 6 AGA	Gln CAG	115 Lys AAG	0 Pro CCC	His CAC	lle AUA	Asp GAU	G I y GGA	Me† AUG	GIy GGC	Thr ACA	G I y GGG	116 Gln CAA	0 Ser AGC	Cys UGC	Trp UGG	lle AUU	Pro CCA	Pro CCA	Ser UCA	Ser AGU	Asp GAC	117 Arg AGA	0 GIy GGA	3866
Pro CCC	Gln CAG	G I u GAA	lle AUA	G I u GAG	G I y GGA	Asn AAC	Ser UCC	118 His CAC	0 Leu CUA	Pro CCC	Ser UCC	Tyr UAC	Arg AGA	Pro CCU	Val GUG	G I y GGG	Pro CCG	119 Glu GAG	0 Lys AAG	Leu CUG	His CAU	Ser UCU	Leu CUC	Gln CAG	Ser UCA	Ala GCU	Asn AAU	120 Gly GGA	0 Ser UCG	3956
Cys UGU	His CAC	Głu GAA	Arg AGG	Ala GCU	Pro CCA	o Asp GAC	Leu CUG	121 Pro CCA	0 Arg CGG	Gln CAG	Thr ACA	Glu GAG	Pro CCA	Val GUU	121 Lys AAG	7 AM UAG	AAA	CUGG	GAGU	AGAC	AGAA	GGUA	CAGA	AGGG	AAAA	UAAC	AAC	CAGG	CUGAUGA	4061
UGG	UAGA	GUGC	UACA	GACU	UGGU	JACUC	CAGU	UUCC	ACGG	CUAA	UCAC	UGCU	CGCU	CAGG	GUCCI	UGAA	GAUA	GCUG	CACA	GCUG	CAGA	GCUG	CACA	GCGG	GAUA	GCUG	CGAC	JUUU	GCUUC	4181
UUG	CUUU	AAGC	AGUU	CCAC	UGAA	GAUA	CUCA	AAAG	AGAA	GUGG	AGAA	AAUC	AUUA	GAAA	CCAA	AGUC	AAGA	CAUU	CAUA	UAUA	AGCU	GUGU	CUUC	JUCA	CUGG	ACGGI	JUUG	CCUC	UUUUC	4301
CUU	UUGC	CUCA	GAAG	GAGU	GGGU	IUAAA	GCAG	guga	cccc	AUGC	UCUG	UCAA	0000	UG <u>AA</u>	UAAA	UGAU	guga	UCUA	CAUA	GAAG	UCUU	AGCU	CACU	CUCA	GGAA	CGCUI	JGGA	ACAC	UAUAA	4421
CUU	UUGC	UAUG	AUAU	ACUG	CCAA	GUGU	IGGCC	CAUG	CUCA	UAAU	UGUG	CCUU	CUGA	AUUG	UGAU.	AAAU	UAGU	GAAA	AAAC	UGUA	ACUU	AGAA	UCUG	AUUU	AUUC	AGGA	JUAG	AUCA	UCUUU	4541
UUA	UACU	AUAA	AAAU	CUUC	GAAU	IGAAA	AUAU	UUAA	CUUU	AAAA	ACAU	UACC	UUAA	UCAU	UGUCI	UUUU	CUUC	UUGA	AGUC	UUUC	CCAG	UGAA	AACG	CUCA	AUUC	UGCU	GUUU	CCAU	AGAAU	4661
UUL	UAAU	UUAU	UUUA	AGAC	AUGA	GAUU	IGUAA	ACAA	AUUG	CUUG	AUUU	AUUU	UAUC	CUAA	UUAU	UUA <u>A</u>	AUAA	AUC	ACCC	UAAA	GCAU	CA								4750

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-4; 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 we determined (14) on both strands and across all restriction sites used to initiate sequence determination.

of eight independently isolated clones indicated that four have Asp¹⁷³ and four have Asn¹⁷³.

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 argininespecific 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 trypsinlike 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

- C. R. Savage, T. Inagani, 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).
 M. Chipkers and S. Cohen. Nature (London)
- 2.
- 43, 251 (1981).
 M. Chinkers and S. Cohen, Nature (London)
 290, 516 (1981); J. E. Kudlow, J. E. Buss, G. N. Gill, *ibid.*, p. 519.
 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). (1978)
- 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).
- J. Scott, M. Selby, M. Urdea, M. Quiroga, G. I. Bell, W. J. Rutter, *Nature (London)* **302**, 538 (1983)
- Oligonucleotides were synthesized by solid-6 phase phosphoramidate methodology [S. L. Beaucage and M. H. Caruthers, *Tetrahedron* Lett. 22, 1859 (1981)]. The sequences in each pool were: pool 1, 3'-CCNCCNCANACApool 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'
- P. S. Thomas, Proc. Natl. Acad. Sci. U.S.A. 77, 7. 5201 (1980).
- Abbreviations for the bases are: A, adenine; C, 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. valine
- 9. D. F. Steiner, P. S. Quinn, S. J. Chan, J. Marsh, D. P. Steiner, P. S. Guinil, S. Chan, S. Chan, S. Marsi, 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).
 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).
 K. A. Thomas, N. C. Baglan, R. A. Bradshaw, *ibid.* 256, 9156 (1981). 11
- 12. Barka, J. Histochem. Cytochem. 28, 836 Τ. (1980)13. Ř
- (1980).
 R. Baserga, Ed., Tissue Growth Factors (Springer-Verlag, New York, 1981); A. B. Rob-erts et al., Nature (London) 295, 417 (1982); L. M. Matrisian, M. Pathak, B. E. Magun, Bio-chem. 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) (1983)
- A. M. Maxam and W. Gilbert, *Methods Enzy-*mol. **65**, 499 (1980); F. Sanger, A. R. Coulson, B. G. Barrell, A. J. H. Smith, B. A. Roe, *J. Mol. Biol.* **143**, 161 (1980).
- 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 15. (J.S.).

17 May 1983