26. G. Steck, P. Leuthard, R. R. Bürk, *Anal. Biochem.* **107**, 21 (1980); T. Rabilloud, G. Carpentier, P. Tarroux, *Electrophoresis* **9**, 288 (1988).
27. Supported by the CNRS, INSERM (CRE-86-4015), the Ligue Nationale Française contre le Cancer, and the Association pour la Recherche contre le Cancer (ARC 6455). C.D. is a fellow of the Ministère de la Recherche et de la Technologie. We thank T. Rabilloud for silver staining; M. Fellous for the gift of antibodies; and C. Steinberg, J. F. Kaufman, B. Bauvois, and J. Gavrilovic for critically reading the manuscript.

27 June 1989; accepted 7 September 1989

Unusual Pattern of Accumulation of mRNA Encoding EGF-Related Protein in Sea Urchin Embryos

QING YANG, LYNNE M. ANGERER, ROBERT C. ANGERER

A sea urchin (*Strongylocentrotus purpuratus*) messenger RNA encoding a protein (SpEGF2) related to epidermal growth factor (EGF) was identified. The full-length complementary DNA sequence predicts a protein with an unusually simple structure, including four tandem EGF-like repeats and a hydrophobic leader, but lacking a potential transmembrane domain. Sequence similarities suggest that the peptides are homologous to two peptides from a different sea urchin species, which cause a classic developmental defect, exogastrulation, when added to the seawater outside of embryos. The SpEGF2 messenger RNA begins to accumulate at blastula stage, and in pluteus larvae it is distributed in discrete regions of ectoderm that are not congruent with known histological borders. One region corresponds to that expressing the homeodomain-containing protein, SpHbox1. The structure of the SpEGF2 protein and the pattern of accumulation of its messenger RNA suggest that it may have important functions as a secreted factor during development of sea urchin embryos.

THE FAMILY OF PROTEINS THAT CONTAIN domains similar to epidermal growth factor (EGF) includes many that are membrane-bound or diffusible mediators of cell-cell interactions. The proteins encoded by the *notch* (*1*) and *delta* (*2*) genes of *Drosophila* and by the *lin-12* (*3*) and *glp-1* (*4*) genes of *Caenorhabditis* are expressed during development of these invertebrate embryos in which they function in decisions of cell fate. The structure of the mRNA encoding a new member of this family, SpEGF2, which is expressed in embryos of the sea urchin *Strongylocentrotus purpuratus*, is shown in Fig. 1A. This message contains an open reading frame of 975 nucleotides beginning with an ATG codon. The deduced protein contains 325 amino acids and has a predicted molecular mass of 36.9 kD. Most (at least 263 of 325 amino acids) of the predicted protein consists of four tandem EGF-like repeats that can be aligned by the six cysteine residues found in each repeat spaced as $CX_6CX_5CX_{8-13}CXCX_{11-12}C$ (Fig. 1B). These are within the range of spacings ($CX_{4-14}CX_{3-8}CX_{8-14}CXCX_{8-14}C$) for 63 EGF-like repeats from functionally diverse proteins (*5*). The SpEGF2 repeats have several highly conserved residues proposed to be important for protein conformation (*6*), including Gly^{18}, Phe/Tyr^{29}, Gly^{36}-Phe/Tyr^{37}, and Gly^{39}. They lack two of six residues, Arg^{41} and Leu^{47}, conserved in ligands that bind to vertebrate EGF receptors (*7*). They also lack the site of β-hydroxylation at Asp/Asn^{22} and several other residues thought to be important for high-affinity Ca^{2+} binding characteristic of some EGF-like repeats of plasma proteins (*5*). Computer-assisted analysis of secondary structure (*8*) predicts that repeats 1 to 3 have a central β-sheet structure similar to that demonstrated for EGF (*9*).

The sequence around the ATG at the beginning of this open reading frame strongly suggests that it is the true initiation codon. First, there are stop codons in all three reading frames upstream of this ATG. Second, the 17 amino acids following it constitute a putative hydrophobic signal peptide, a feature common to all EGF-related proteins. Third, primer extension identified the A residue 70 base pairs upstream of this ATG codon as the unique transcription initiation site. Furthermore, sequence upstream from this initiation site includes a TATA box at base −32 to −26 and (G)CCAATT motifs at bases −52 and −171. Several other close matches to known regulatory motifs, whose functional significance in the SpEGF2 gene remains to be established, are described in the legend to Fig. 1A. The 3′ untranslated region is approximately 500 bases long and contains a known variant signal for transcript cleavage and polyadenylation, ATTAAA (*10*), which begins 19 bases upstream from a stretch of A residues that presumably represent the beginning of the polyadenylate tail. These features predict an mRNA length which matches that determined by blotting analysis.

The EGF-like domains ultimately have extracellular destinations, either as secreted peptides or in extracellular regions of membrane-bound proteins. In addition to their initially identified roles as growth factors [for example, EGF and transforming growth factor–α (TGF-α)], EGF-like domains are found in mosaic proteins that function in diverse biological processes and usually contain various other functional domains. However, several lines of evidence suggest that the SpEGF2 mRNA encodes a protein of relatively simple structure that is secreted into the blastocoel and whose function is mediated largely, if not entirely, by the EGF-like domains. (i) Suyemitsu *et al.* (*11*) recently reported the sequences of two of four peptides found in the blastocoel of embryos of a different sea urchin, *Anthocidaris crassispina*, which are similar to the SpEGF2 repeats (Fig. 1B). The similarity is greatest for peptide A, which has about 60% identity to repeats 2 and 3 of SpEGF2 and 64% to 68% similarity when conservative amino acid substitutions are included. This is a strong similarity in view of the time since divergence of the lines leading to these two sea urchin species [30 to 45 million years ago (*12*)]. EGF and TGF-α are only 55% identical in amino acid sequence, but bind to the same receptor with similar affinity (*7*). Furthermore, the peptides from *A. crassispina* have the same developmental effects on embryos of both genera. In contrast to their similarity with peptides of *A. crassispina*, the SpEGF2 repeats differ from the other EGF-related protein identified in *S. purpuratus*, uEGF1, which contains at least 20 relatively precise repeats of the EGF-like domain as well as other functional domains (*13*). Thus, it seems likely that *A. crassispina* peptides and SpEGF2 protein encode functionally equivalent products of homologous genes. (ii) The SpEGF2 protein includes a probable leader peptide but lacks a sequence suitable to serve as a transmembrane domain. (iii) At late blastula stage, SpEGF2 mRNA is concentrated at the basal sides of cells (*14*), a phenomenon we have observed for only one other mRNA, one that encodes an extracellular arylsulfatase (*15*). (iv) Other than the four EGF-like repeats and hydrophobic leader, the protein contains two stretches of 30 or fewer amino acids surrounding the EGF-like repeats, which do not correspond to any sequence in the National Biomedical Research Foundation protein data base. (v) Several other features

Department of Biology, University of Rochester, Rochester, NY 14627.

of the sequence suggest that the SpEGF2 protein may be cleaved and modified: repeats 1 to 4 are preceded by Arg-Asp at a conserved position, and the four related peptides from *A. crassispina* all have NH_2-terminal Asp residues (*11*). Three pairs of basic amino acids, two of which flank the fourth repeat, could also serve as proteolytic processing sites (*16*). The predicted peptide lacks N-X-S/T signals for N-linked glycosylation (*17*) but contains several clusters of serine or threonine (or both) that could serve as sites of O-linked glycosylation.

Hybridization in situ showed that SpEGF2 mRNA accumulates only in ectoderm in a spatial pattern that is unusual because it does not correspond to known histological borders. In pluteus larvae (Fig. 2) a major site of SpEGF2 mRNA accumulation is in the histologically uniform cells of aboral ectoderm. In contrast to nine other messages expressed in aboral ectoderm (*18*), SpEGF2 transcripts are not uniformly distributed in this tissue but instead accumulate to higher levels at the pointed vertex opposite the mouth and at the border with the ciliary band. This nonuniform distribution, and the manner in which it is progressively established from uniform distribution in presumptive aboral ectoderm of the blastula (*14*), are reminiscent of similar phenomena observed for another mRNA, SpHbox1. The SpHbox1 mRNA encodes a putative transcription factor bearing a homeodomain and becomes progressively restricted during embryogenesis to aboral ectoderm cells at the vertex of the pluteus larva (*19*). However, the tissue specificity of the SpEGF2 message is broader than that of SpHbox1 mRNA in that it also is expressed in oral ectoderm cells (Fig. 2). In these cells, also, it is inhomogenously distributed among epithelial cells that surround the mouth, being most abundant in cells immediately adjacent to the mouth on its anal side, and absent from the histologically distinct cells that constitute the ciliary band.

Although these in situ hybridizations show a complex pattern of SpEGF2 mRNA accumulation, they monitor the distribution of a single 1.8-kb mRNA, as shown by blotting analysis (Fig. 3) that was carried out at comparable stringency. The SpEGF2 mRNA is undetectable in unfertilized eggs, begins to accumulate around mesenchyme blastula stage to an appreciable fraction of its maximum abundance, which occurs during gastrulation, and persists at slightly reduced levels in pluteus larvae. This time course is typical of tissue-specific mRNAs that accumulate in differentiating aboral ectoderm, which is a major site of expression of SpEGF2.

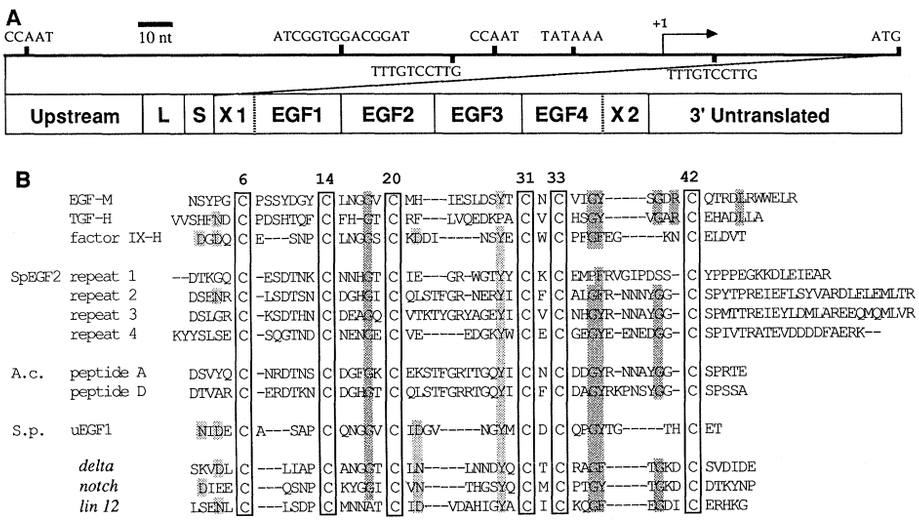

Fig. 1. (**A**) Structure of SpEGF2 mRNA and upstream sequence motifs. The SpEGF2 message contains a 70-nucleotide (nt) 5′ untranslated leader (L), a putative hydrophobic signal peptide (S), four tandem EGF-like repeats (EGF1 to EGF4), two short stretches of unidentified amino acid sequence whose borders with the EGF-like repeats cannot be precisely determined (X1 and X2), and a 3′ untranslated sequence of approximately 500 bases (3′ untranslated). Sequence motifs upstream of the open reading frame include two 10–base pair perfect direct repeats, TTTGTCCTTG, centered at bases −63 and 17, which contain the motif most strongly conserved within the glucocorticoid response element [TGT(T/C)CT] (*20*), and ATCGGTGGACGGAT centered at base −102, which matches 12 of 14 nucleotides in an inverted repeat of the bovine papillomavirus enhancer (*21*). The original SpEGF2 cDNA clone was isolated by a differential screen for mRNAs that increase in abundance during development (*15*), and longer cDNA clones were subsequently identified in a pluteus cDNA library in λgt11. Inserts were subcloned into M13, and nucleotide sequences were determined as described (*15*). DNA sequences and deduced amino acid sequences were analyzed by the computer programs developed by Pustell and Kafatos (*22*). The transcription start site (+1) was determined as described (*15*) with 1 ng of a synthetic 21-base oligonucleotide corresponding to nucleotides 310 to 330 hybridized overnight with 35 μg of total gastrula RNA. (**B**) Amino acid sequences of SpEGF2 and other EGF-like repeats. The amino acid sequences of the four EGF-like repeats of SpEGF2 were deduced from the complete cDNA sequence. The other sequences shown are: A. c., exogastrulation-inducing peptides A and D from *A. crassispina* (*11*); uEGF1, one of approximately 20 EGF-like repeats in the other EGF-related gene of *S. purpuratus* (S.p.), which most closely matches the consensus of the nine repeats whose sequences have been reported (*13*); EGF-M, mouse EGF (*23*); TGF-H, human transforming growth factor–α (*24*); factor IX-H, the first repeat of human clotting factor IX (*25*); *notch* and *delta*, the fifth repeat and eighth (complete) repeat, respectively, in these genes of *Drosophila* (*1, 2*); and *lin-12*, the ninth repeat in open reading frame A of the *lin-12* gene of *Caenorhabditis elegans* (*3*). The sequences are aligned on the six cysteine residues (outlined) and numbered according to positions in mouse EGF. Strongly conserved residues, including six found in peptides that bind to vertebrate EGF receptors (EGF-M and TGF-H) are highlighted by dark shading. Other residues conserved in members of the family that bind Ca^{2+} (for example, factor IX-H) are highlighted by light shading. Abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

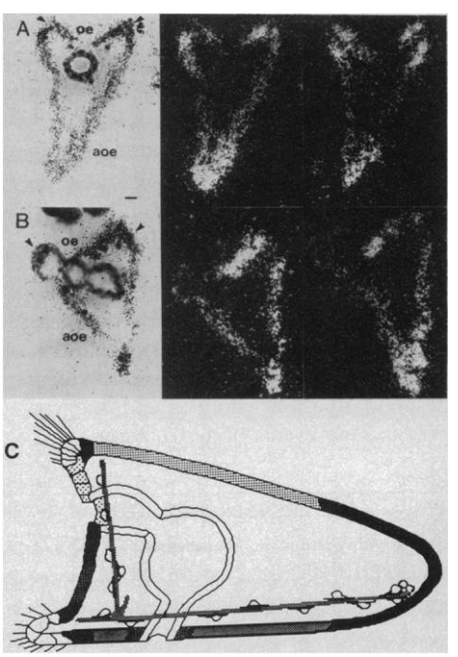

Fig. 2. Distribution of SpEGF2 mRNA in pluteus larvae. Each row shows two adjacent sections photographed in dark-field illumination, and one of these photographed in phase-contrast. (**A**) Sections cut parallel to the anal side and passing through the intestine and anal arms. The oral side of the embryo is at the top. (**B**) Sections corresponding to the diagram shown in (C). Arrowheads point to unlabeled cells of the ciliary band. Abbreviations: oe, oral ectoderm; aoe, aboral ectoderm. (**C**) Diagrammatic representation of the labeling pattern, in which darker shading indicates higher SpEGF2 message concentration. Hybridization in situ with ^{35}S-labeled antisense RNA probes (5×10^8 dpm/μg) was carried out as described (*26*). Autoradiographic exposures were for 1 week. Scale bar in (A), 10 μm.

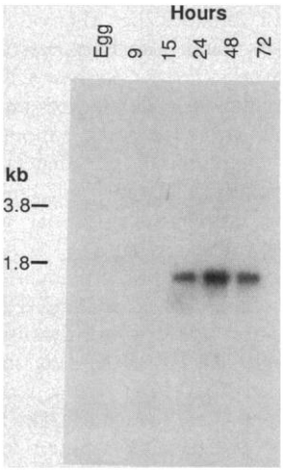

Fig. 3. Accumulation of SpEGF2 mRNA during development. Two micrograms of total RNAs from eggs and embryos at the indicated times of development were analyzed by blot hybridization as described (*15*) with a probe corresponding to the sequence of amino acid 303 to the 3′ end of the mRNA. Exposure was for 1 day with two intensifying screens. Developmental times of 9, 15, 24, 48, and 72 hours correspond to late cleavage, early blastula, mesenchyme blastula, late gastrula, and pluteus stages, respectively. Size calibration was provided by positions of the ribosomal RNAs.

The exact developmental role of the SpEGF2 peptide (or peptides) is unknown. The EGF-like peptides present in the blastocoelic fluid of *A. crassispina* cause exogastrulation of embryos of several different sea urchin species, including a species of *Strongylocentrotus*, when added to the outside of embryos (*11*). Their immediate morphological effect on blastulae, thinning of the vegetal plate, implies that precursor cells of endoderm and secondary mesenchyme can respond to these peptides and suggests a role in early events of gastrulation.

REFERENCES AND NOTES

1. K. A. Wharton, K. M. Johansen, T. Xu, S. Artavanis-Tsakonas, *Cell* **43**, 567 (1985).
2. H. Vässin *et al.*, *EMBO J.* **6**, 3431 (1987).
3. I. Greenwald, *Cell* **43**, 583 (1985).
4. J. Yochem and I. Greenwald, *ibid.* **58**, 553 (1989); J. Austin and J. Kimble, *ibid.*, p. 565.
5. D. J. G. Rees *et al.*, *EMBO J.* **7**, 2053 (1988).
6. H. Marquardt, M. W. Hunkapiller, L. E. Hood, G. J. Todaro, *Science* **223**, 1079 (1984).
7. A. W. Burgess, *Br. Med. Bull.* **45**, 401 (1989).
8. P. Y. Chou and G. D. Fasman, *Adv. Enzymol.* **47**, 45 (1978); J. Garnier, D J. Osguthorpe, B. Robson, *J. Mol. Biol.* **120**, 97 (1978).
9. K. H. Mayo, P. Schaudies, C. R. Savage, A. De Marco, R. Kapstein, *Biochem. J.* **239**, 13 (1986); G. T. Montelione, K. Wüthrich, E. C. Nice, A. W. Burgess, H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 5226 (1987); K. Makino *et al.*, *ibid.*, p. 7841; R. M. Cooke *et al.*, *Nature* **327**, 339 (1987).
10. M. L. Birnstiel *et al.*, *Cell* **41**, 349 (1985).
11. T. Suyemitsu *et al.*, *Cell Diff. Dev.* **26**, 53 (1989).
12. A. B. Smith, *Mol. Biol. Evol.* **5**, 345 (1988).
13. D. A. Hursh, M. E. Andrews, R. A. Raff, *Science* **237**, 1487 (1987).
14. Q. Yang, L. M. Angerer, R. C. Angerer, unpublished observations.
15. ———, *Dev. Biol.* **135**, 53 (1989).
16. D. D. Sabatini, G. Kreibich, T. Morimoto, M. Adesnik, *J. Cell Biol.* **92**, 1 (1982).
17. W. C. Barker, L. T. Hunt, D. G. George, *Protein Seq. Data Anal.* **1**, 363 (1988).
18. R. C. Angerer *et al.*, in *Society for Experimental Biology Seminar Series*, N. Harris and D. Wilkinson, Eds. (Cambridge Univ. Press, Cambridge, in press).
19. L. M. Angerer *et al.*, *Genes Dev.* **2**, 239 (1988).
20. H.-M. Jantzen *et al.*, *Cell* **49**, 29 (1987).
21. C. Moskaluk and D. Bastia, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 1215 (1987).
22. J. Pustell and F. C. Kafatos, *Nucleic Acids Res.* **10**, 51 (1982); *ibid.* **12**, 643 (1984).
23. A. Gray, T. J. Dull, A. Ullrich, *Nature* **303**, 722 (1983); J. Scott *et al.*, *Science* **221**, 236 (1983).
24. R. Derynck, A. B. Roberts, M. E. Winkler, E. Y. Chen, D. V. Goeddel, *Cell* **38**, 287 (1984).
25. K. Kurachi and E. W. Davie, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 6461 (1982).
26. L. M. Angerer, K. H. Cox, R. C. Angerer, *Methods Enzymol.* **152**, 649 (1987).

27. Supported by NIH grant GM25553 to R.C.A. and L.M.A. R.C.A. is the recipient of a Career Development Award from USPHS (HD602). We thank E. Davidson for the pluteus cDNA library; S. Reynolds for providing RNA for determination of the transcription start site by primer extension; New England Nuclear for ^{35}S-labeled uridine 5′-triphosphate; and M. Gorovsky, P. Hinkle, V. Stathopoulos, and Y. Xiong for helpful comments on the manuscript. The SpEGF2 cDNA sequence has been submitted to GenBank and assigned the accession number m29004.

31 May 1989; accepted 15 September 1989

Chromosomal Location and Evolutionary Rate Variation in Enterobacterial Genes

PAUL M. SHARP, DENIS C. SHIELDS, KENNETH H. WOLFE, WEN-HSIUNG LI

The basal rate of DNA sequence evolution in enterobacteria, as seen in the extent of divergence between *Escherichia coli* and *Salmonella typhimurium*, varies greatly among genes, even when only "silent" sites are considered. The degree of divergence is clearly related to the level of gene expression, reflecting constraints on synonymous codon choice. However, where this constraint is weak, among genes not expressed at high levels, divergence is also related to the chromosomal location of the gene; it appears that genes furthest away from *oriC*, the origin of replication, have a mutation rate approximately two times that of genes near *oriC*.

EVOLUTIONARY RATES AT THE DNA sequence level are determined by the nucleotide mutation rate and the subsequent effects of natural selection; generally, only the latter is considered to vary among genes in a genome. If "silent" codon positions (that is, those which can be changed without affecting the amino acid sequence of the gene product) are effectively neutral as suggested (*1*), the nucleotide substitution rate at silent sites should reflect only the mutation rate (*2*). Thus, synonymous nucleotide substitutions should accumulate at similar rates in different genes and potentially constitute a useful molecular clock (*3*). However, silent substitution rates have been found to vary among genes in animals (*4, 5*) and bacteria (*6, 7*); the question is whether this reflects variation in mutation rates, selective constraints, or both.

In *Escherichia coli* it is clear that synonymous codons are not selectively equivalent: codon usage is highly biased, the "preferred" codons are those recognized most accurately and/or efficiently by the most abundant tRNA species, and the degree of bias is greater in genes expressed at high levels (*8*). Both tRNA levels (*9*) and patterns of codon usage (*7, 9*) are conserved between *E. coli* and *Salmonella typhimurium*, and so it would be expected that genes retaining high codon usage bias in both species must have accumulated fewer substitutions. This is indeed the observation: comparison of 67 pairs of homologous gene sequences (*10*) from the two species (Fig. 1A) shows that the extent of divergence at silent sites (expressed as the corrected number of base substitutions per synonymous site, K_S) is significantly negatively correlated with the degree of bias in codon usage (Table 1, model 1). (Codon usage bias is measured by the codon adaptation index, CAI; see legend to Fig. 1.) Thus, selection among synonymous codons, which is most prevalent in highly expressed genes, forms an evolutionary constraint (*7*). However, this cannot explain all of the variation in degree of divergence; in particular there are some genes with low codon bias but surprisingly low divergence (such as those indicated by open circles in Fig. 1A).

We noticed that these anomalous (that is, having low codon bias and low divergence) genes are located near *oriC*, the origin of replication of the bacterial chromosome at

P. M. Sharp, D. C. Shields, K. H. Wolfe, Department of Genetics, Trinity College, Dublin 2, Ireland.
W.-H. Li, Center for Demographic and Population Genetics, University of Texas, Houston, TX 77030.