THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 1990 by The American Society for Biochemistry and Molecular Biology, Inc.
Vol. 265, No. 14, Issue of May 15, pp. 7872-7879, 1990
Printed in U.S.A.

cDNA and Gene Analyses Imply a Novel Structure for a Rat Carcinoembryonic Antigen-related Protein*

(Received for publication, October 16, 1989)

Sabine Rebstock, Kurt Lucas, John A. Thompson, and Wolfgang Zimmermann‡

From the Institut für Immunbiologie, Universität Freiburg, Stefan-Meier-Str. 8, D-7800 Freiburg, Federal Republic of Germany

The gene encoding the human tumor marker carcinoembryonic antigen (CEA) belongs to a gene family which can be subdivided into the CEA and the pregnancy-specific glycoprotein subgroups. The corresponding proteins are members of the immunoglobulin superfamily, characterized through the presence of one IgV-like domain and a varying number of IgC-like domains. Since the function of the CEA family is not well understood, we decided to establish an animal model in the rat to study its tissue-specific and developmental stage-dependent expression. To this end, we have screened an 18-day rat placenta cDNA library with a recently isolated fragment of a rat CEA-related gene. Two overlapping clones containing the complete coding region for a putative 709 amino acid protein (rnCGM1; M_r = 78,310) have been characterized. In contrast to all members of the human CEA family, this rat CEA-related protein consists of five IgV-like domains and only one IgC-like domain. This novel structure, which has been confirmed at the genomic level might have important functional implications. Due to the rapid evolutionary divergence of the rat and human CEA gene families it is not possible to assign rnCGM1 to its human counterpart. However, the predominant expression of the rnCGM1 gene in the placenta suggests that it could be analogous to one of the human pregnancy-specific glycoprotein genes.

The human tumor marker carcinoembryonic antigen (CEA)[1] is a large cell surface glycoprotein with a relative molecular mass (M_r) of 180,000. Rising CEA serum concentrations are widely used as an indicator of tumor recurrencies during the postoperative surveillance, especially of patients with colorectal carcinomas (Shively and Beatty, 1985).

CEA is a member of a family of cross-reacting proteins encoded by approximately 14 genes in man (Thompson, *et al.*, 1989),[2] which can be grouped together in the CEA and pregnancy-specific glycoprotein (PSG) subgroups based on sequence comparisons. PSGs are produced in large amounts during pregnancy in the placenta and secreted into the maternal serum (Lin *et al.*, 1974). Furthermore, these proteins have been discussed as potential markers for trophoblastic tumors because of their presence in choriocarcinomas (Tatarinov, 1978). The members of the human CEA family belong to the immunoglobulin (Ig) super family and are composed of one N-terminal or Ig variable (IgV)-like domain and a varying number of half-repeats (R_A,R_B) or Ig constant (IgC)-like domains (Thompson *et al.*, 1989a). The *in vivo* functions of the CEA/PSG-like proteins during fetal development and in neoplasia are not clear, although recent findings by Benchimol *et al.* (1989) indicate a role in cell adhesion for CEA.

In this context it would be important to analyze the temporal and spatial expression pattern of the CEA gene family. In order to overcome the problem of limited availability of human tissues and to study expression pattern during embryogenesis, we and others have started to establish animal models. Recently, Beauchemin *et al.* (1989) identified a CEA-related cDNA derived from mouse colon RNA. Furthermore, we could show that in rat a gene family exists, the members of which show obvious homology but a relatively low sequence similarity to the human CEA-like genes (Kodelja *et al.*, 1989). Sequence and Southern blot analyses indicate that the CEA-related genes in primates and rodents must have evolved independently but in parallel, from one or only a few common ancestral genes (Rudert *et al.*, 1989). For this reason, it is presently not possible to assign individual rodent genes to their human counterparts.

To be able to study the function of CEA and related proteins in a rodent model, further information on this rapidly evolving gene family is needed. To this end we have isolated several clones from a rat placental cDNA library using a genomic probe that covers an N-terminal domain exon of the CEA-related rat gene rnCGM1 (Kodelja *et al.*, 1989). Sequence determination of overlapping cDNA clones have revealed a novel domain organization for the encoded rat CEA-related protein, which has been confirmed by sequence analysis of the corresponding gene. Northern analyses have demonstrated that this gene is expressed in placenta but not in a panel of other fetal or adult tissues.

MATERIALS AND METHODS

Isolation of Rat cDNA Clones—Phages from a λgt10 cDNA library (complexity 1×10^7 independent clones) constructed from RNA of an 18-day rat placenta (a generous gift from M. L. Duckworth, The University of Manitoba Faculty of Medicine, Winnipeg) were plated onto *Escherichia coli* MA 150 (Young and Davis, 1983) and transferred to nitrocellulose membranes in replicas. The phage DNAs were hybridized with the ^{32}P-labeled (Feinberg and Vogelstein, 1983) 1.1-kb *Eco*RI fragment from the rat genomic clone rnCGM1-1 as described before (Kodelja *et al.*, 1989). The filters were first washed under low stringency (2 × SSPE, 60 °C (1 × SSPE: 180 mM NaCl, 10 mM sodium phosphate, pH 7.4, 1 mM EDTA)), and after autoradiography, rewashed under more stringent conditions (0.5 × SSPE, 65 °C). Clones which were further analyzed were plaque purified twice.

Sequence Determination and Analysis—Suitable cDNA fragments

* This work was supported by a grant from the Dr. Mildred Scheel Stiftung für Krebsforschung. 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.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EMBL Data Bank with accession number(s) J05417.

‡ To whom reprint requests should be addressed.

[1] The abbreviations used are: CEA, carcinoembryonic antigen; PSG, pregnancy-specific glycoprotein; kb, kilobase; bp, base pair; SDS, sodium dodecyl sulfate.

[2] J. A. Thompson, unpublished results.

were subcloned in M13mp18/mp19 or the phagemid Bluescript (Stratagene, La Jolla, California). For sequence determination of the cDNA fragments, subclones of opposite orientation were identified by the C-test (Messing *et al.*, 1983) so that the sequence of both strands could be determined. Sequencing was performed according to Sanger *et al.* (1977) on single or double-stranded templates using internal or universal oligonucleotide primers. Internal primers were synthesized according to the manufacturers protocol (Applied Biosystems, Weiterstadt, Federal Republic of Germany (F. R. G.)) and purified as described (Kodelja *et al.*, 1989). For calculation of similarities at the nucleotide and amino acid level, the computer program "align" was used.[3] To determine the similarity, pairs of amino acids with logarithms of odd matrix scores ≥9 were taken as conservative exchanges (Feng *et al.*, 1985). Further sequence analyses were carried out using the program package PCgene (Genofit, Heidelberg, F. R. G.). The frequencies of synonymous substitutions were calculated using the computer program LWL85 (Li *et al.*, 1985). The nucleotide sequences of this paper have been submitted to GenBank (Los Alamos).

Mapping of the Genomic Clone λrnCGM1-1—The isolation of the EMBL3 λ phage rnCGM1-1 which contains parts of the rat gene rnCGM1 has been previously described (Kodelja *et al.*, 1989). For mapping of restriction endonuclease sites, *Kpn*I and *Nhe*I were used in separate experiments to cut the left or right arm of the EMBL vector at a distance of 1221 and 180 bp, respectively, from the cloning site. Subsequently, the recombinant phage DNA was partially digested with either *Bam*HI, *Eco*RI, or *Sal*I (3 units/μg DNA) for 1, 2.5, 5, or 60 min. After size separation on a 0.6% agarose gel the DNA was blotted onto a charged nylon membrane and hybridized separately with ^{32}P-labeled oligonucleotides, complementary to λ sequences flanking the cloning site:

oligo EMBL3-L: 5′GAGTCTTGCAGACAAACTGCGCAAC 3′

oligo EMBL3-R: 5′AGGTTCATTACTGAACACTCGTCCG 3′.

Hybridization was carried out overnight at 35 °C in the presence of 1 M NaCl, 30% formamide, 10% dextran sulfate, 1% SDS. The blots were washed twice for 15 min at 53 °C in 2 × SSPE, 0.5% SDS. For the identification of exon containing fragments, overlapping restriction fragments were generated by single and double digests using *Eco*RI, *Bam*HI, and *Sal*I, size fractionated, blotted, and hybridized with the ^{32}P-labeled cDNA insert of λrnCGM1b.

Primer Extension Experiment—A 23-mer oligonucleotide (5′GGGAGTACACCTCTTGCAGGGAAG 3′), complementary to position 18–41 of the cDNA, was 5′-end labeled with [^{32}P]dATP using T_4 polynucleotide kinase. For the annealing reaction 4 ng of the ^{32}P-labeled oligonucleotide was mixed with 2.5 μg of poly(A) RNA isolated from placenta at day 16 of gestation in 10 μl of 5 mM sodium phosphate buffer, pH 7.0, 5 mM EDTA. After denaturation at 90 °C for 5 min, NaCl was added to a final concentration of 80 mM, and hybridization was carried out at 50 °C for 1 h. Then the mixture was allowed to cool slowly to room temperature. The hybridization mix was adjusted to 17.5 mM Tris/Cl, pH 8.3, 4.3 mM $MgCl_2$, 1.75 mM dithiothreitol, 3.5 mM dNTP, 1 ng/μl actinomycin D, 2 units/μl RNasin (Atlanta, Heidelberg, F. R. G.), 0.8 units/μl avian myeloblastosis virus reverse transcriptase (Boeringer Mannheim, Mannheim, F. R. G.) in 25 μl of volume and incubated at 42 °C for 1 h. The reaction was stopped by phenol extraction and ethanol precipitation. The extension products were denatured and analyzed on a 6% polyacrylamide DNA sequencing gel, containing 8 M urea.

RNA Isolation and Northern Blot Analyses—Tissues were taken from BDHII rats (Druckrey, 1971) that had been anesthetized and killed by cervical dislocation and immediately frozen in liquid nitrogen. Total RNA was isolated by the guanidine thiocyanate method (Fiddes and Goodman, 1979). Poly(A) RNA was isolated by one round of chromatography on oligo(dT)-cellulose (Sigma, Deisenhofen, F. R. G.) according to Aviv and Leder (1972). For Northern blot hybridization, 2 μg of poly(A) RNA were size fractionated on a 1% agarose gel containing 10 mM methylmercury hydroxide and transferred onto a charged nylon membrane (GeneScreen-Plus, Du Pont de Nemours, Bad Homburg, F. R. G.) (Alwine *et al.*, 1977). Alternatively, 10 μg of total RNA were size separated on a 1% agarose gel containing 2.2 M formaldehyde and blotted onto GeneScreen-Plus membranes according to the manufacturer's protocol. Hybridization with ^{32}P-labeled cDNA fragments was carried out at 42 °C in the presence of 1 M NaCl, 10% dextran sulfate, 40% formamide overnight. Final washes were carried out at 60 °C in the presence of 2 × SSPE, 0.1% SDS or at 65 °C in 0.1 × SSPE, 0.1% SDS.

RESULTS

Isolation and Analysis of Rat CEA-related cDNA Clones—Since we had previously shown that members of the rat CEA gene family are strongly expressed in placenta (Kodelja *et al.*, 1989), we screened a cDNA library which had been constructed from RNA of an 18-day rat placenta. As a probe we used the 1.1-kb *Eco*RI fragment of the genomic clone λrnCGM1-1, that contains an N-terminal domain exon (Kodelja *et al.*, 1989). Hybridization of 2.5 × 10^5 plaques under low stringency conditions yielded about 120 positive clones. We raised the wash stringency stepwise and could distinguish groups of clones that showed different degrees of sequence similarity to rnCGM1. Three clones were further analyzed, one of which showed a reduced hybridization signal after the highest stringency wash and revealed identity to the analyzed N-terminal domain exon of rnCGM3 (Kodelja *et al.*, 1989). This finding proves that the corresponding gene is actively transcribed. The inserts of the two other clones, λrnCGM1a and λrnCGM1b were found to overlap. Their nucleotide and deduced amino acid sequence, as well as their sequencing strategy are shown in Fig. 1. The sequences of the two cDNA clones rnCGM1a/1b are identical in the overlapping region (position 219–759) with the exception of position 283 where λrnCGM1a contains an A, λrnCGM1b and the genomic clone rnCGM1-1 a G. This would lead to an amino acid change in the encoded putative protein (Arg to His). The entire cDNA has a length of 3190 nucleotides and covers a 121-bp 5′-noncoding sequence, an open reading frame of 2127 bp, corresponding to 709 amino acids (M_r = 78,310, M_r minus leader = 74,553) and a 3′-noncoding sequence of 942 bp (Fig. 2). The 3′-noncoding sequence includes a so-called simple sequence [$(CCTT)_8$]. Simple sequences are widely distributed in the genome of eucaryotes and are found in many genes, *e.g.* in the first intron of rnCGM1 (Kodelja *et al.*, 1989). Since no polyadenylation consensus sequence could be found, the actual length of the 3′-noncoding sequence could not be assessed. Sequence comparisons with partial sequences of rat CEA-like genes revealed that the two cDNA clones correspond to the

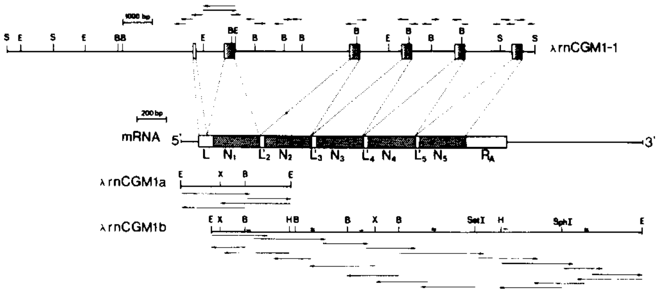


FIG. 1. **Structure and sequencing strategy of the rnCGM1 gene and cDNAs.** The upper part of the figure shows the restriction map and exon arrangement of the genomic clone λrnCGM1-1, the lower part, the domain structure of the cDNA clones λrnCGM1a/1b. Exons and domains are depicted as *blocks*, whereby homologous coding regions are indicated by the same *shading*. The corresponding mRNA and genomic regions are connected by *dotted lines*. The coding regions are composed of a leader peptide (*L*), five N-terminal or IgV-like domains (N_1–N_5), four interspersed leader-like sequences (L'_2–L'_5) and an IgC-like or type-A half-repeat (R_A). Introns as well as noncoding sequences are indicated as *lines*. Restriction endonuclease sites used for mapping and subcloning are indicated: *S* = *Sal*I, *E* = *Eco*RI, *B* = *Bam*HI, *X* = *Xmn*I, *H* = *Hpa*II. The sequencing strategy is shown by *arrows*; *small arrows* above the clones λrnCGM1a/b represent synthetic oligonucleotide primers used for determination of internal sequences.

[3] M. Trippel and R. Friedrich, unpublished results.

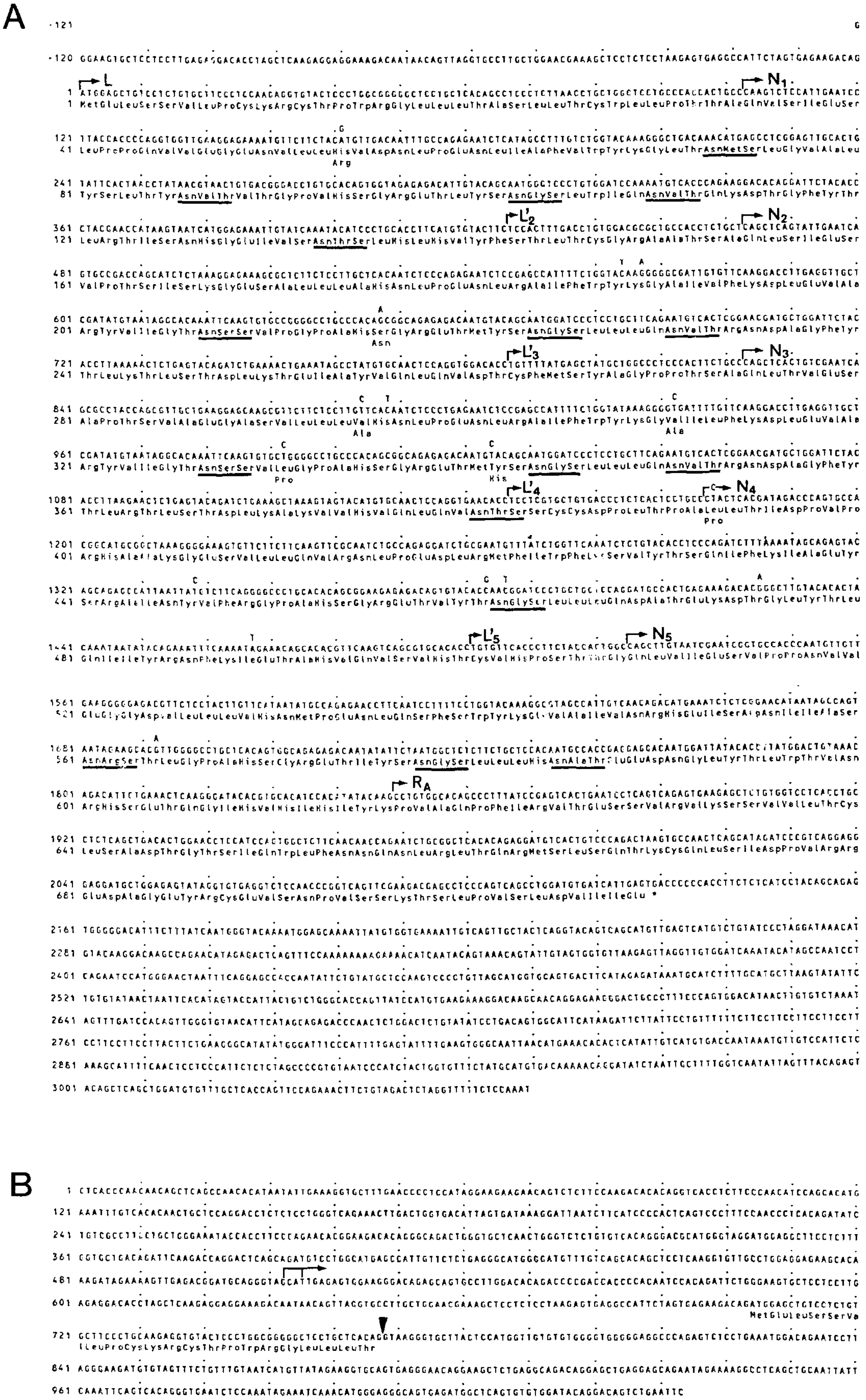
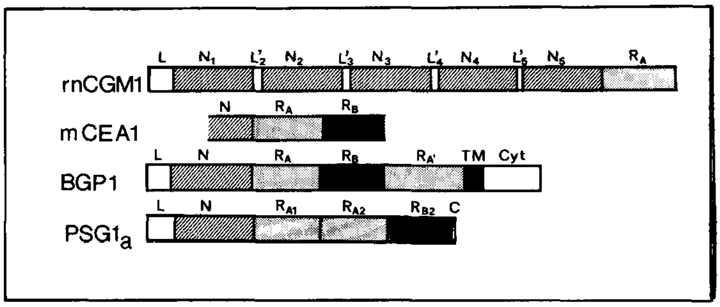


FIG. 2. **Nucleotide sequence of the cDNA and the putative promoter region of the rnCGM1 gene.** *A,* nucleotide and amino acid sequences of the rnCGM1 cDNA. The domain borders are indicated by *arrows.* For abbreviations see legend to Fig. 1. The recognition sequences for potential *N*-glycosylation are *underlined.* Nucleotides and corresponding amino acid sequence differences found in the genomic clone λrnCGM1-1 are shown above and below the main sequence, respectively. *B,* nucleotide sequence of the 1053-bp *Eco*RI fragment of

Leader

```
rnCGM1   L      MELSSVLPCKRCTPWRGLLLTASLLTCWLLPTTA
         L'2                           STLTCGRAA*SA
         L'3                           CFMSYAGPP*SA
         L'4                            SSCCDPL*PA
         L'5                              CVHPS*TG
BGP1            *GHL*APLHRVRVP*QG*****SLLTFWNPP*TA
PSG1            *GTL*APPCTQRIK*KG*****SLLNFWNLP*TA
```



N-terminal domain (IgV-like domain)

```
rnCGM1     QVSIESLPPQVVEGEN.VLLHVDNLPENLIAFV.WYKGL..TNMSLGVALYSLTYNVTVTGPVHSGRETLYSNGS..LWIQNVTQKDTGFYTLRTISNHGEIVSNTSLHLHVYF
       N2  QLSIESV*TSISK*ES.A**LAH*L*ENLRAIF.*Y*GA.IVFKDLEVARYVIGTNSSVP*PAH****TM*S*G*..*LLQNV*RN*A*F***KTLSTDLKTEIAYVQLQVDT
       N3  QLTVESA*TSVAE*AS.V**LVH*L*ENLRAIF.*Y*GV.ILFKDLEVARYVIGTNSSVL*PAH****TM*S*G*..*LLQNV*RN*A*F***RTLSTDLKAKVVHVQLQVNT
       N4  LLTIDPV*RHAAK*ES.V**QVR*L*EDLRMFI.*F*SV.YTSQIFEIAEYSRAINYVFR*PAH****TV*T*G*..*LLQDA*EK*T*L***QIIYRNFKIETAHVQVSVHT
       N5  QLVIESV*PNVVE*GD.V**LVH*M*ENLQSFS.*Y*GV.AIVNRHEISRNIIASNRSTL*PAH****TI*S*G*..*LLHNA*EE*N*L***WTVNRHSETQGIHVHIHIYK
mCEA1                                                   PQFVPNSNMNFT*QAY****II*S*G*..*LFQMI*MK*M*V***DMTDENYRRTQATVRFHVH
BGP1       QLTTESM*FNVAE*KE.V**LVH*L*QQLFGYS.*Y*GERVDGNRQIVGYAIGTQQAT.P*PAN****TI*P*A*..*LIQNV*QN*T*F***QVIKSDLVNEEATGQFHVYP
PSG1       QVTIEAE*TKVSE*KD.V**LVH*L*QNLTGYI.*Y*GQMRDLYHYITSYVVDGEIII.Y*PAY****TA*S*A*..*LIQNV*RE*A*S***HIIKGDDGTRGVTGRFTFTLH
```

half repeat A (IgC-like domain)

```
rnCGM1      KPVAQPFIRVTESSVRVKSS.VVLTCLSADTGTS..IQWLFNN..QNL..................RLTQRNSL.SQTKCQLSIDPVRREDAGEYRCEVSNPVSSKTSLPVSLDVIIE
mCEA1       PILLK*N*TSNNSNPVEGDDS*S***DSYTDPDNINYL*SRNG..ES*..................SEGDRLK*.*EGNRT*TLLN*TRNDT*P*V**TR*P**VNR*D*FS*NIIY
PSG1 A1     LGTPK*S*SSSNLNPRETMEA*S***DPETPDAS..YL*WMNG..QS*..................PMTHSLK*.*ETNRT*FLLG*TKYTA*P*E**IR*P**ASR*D*VT*NLL
BGPI A'    SPVVAK*Q*KASKTTVTGDKDS*N***STNDTGIS..IR*FFKN..QS*..................PSSERMK*.*QGNTT*SINP*KREDA*T*W**VF*PI*KNQ*D*IM*NVNY
```

half repeat B (**IgC-like domain**)

```
mCEA1       GPDTPIISPSDIYLHPGSN.LNLSCHAASNPPAQ.YFWLINEKPN.............................ASSQELFIPNITTNNSGTYTCFVNNSVTGLSRTTVKNITVL
BGP1        ***T*T*S**DT*YRP*AN.*S***Y*A******.*S*L**GTFQ.............................QST*E***PN**VN***S*T*HAN**V**CNRTTVKTII*TEL
PSG1 B2     ***L*R*Y**FT*YRS*EV.*Y***S*D******.*S*T**EKFQ.............................LPG*K***RN**TKP**L*V*SVR**A**KESSKSMTVE*S
```

FIG. 3. **Alignment of amino acid sequences of human and rodent CEA-like proteins.** The deduced amino acid sequences in *one-letter* code of the leader, N-terminal, and half-repeat domains from the following human and rodent CEA-like proteins have been aligned: *BGP1*, human biliary glycoprotein 1 (Hinoda *et al.*, 1988); *PSG1*, human pregnancy-specific glycoprotein 1 (Zimmermann *et al.*, 1989), *mCEA1*, murine CEA-like protein (Beauchemin *et al.*, 1989) and rat rnCGM1. Within a domain, conservatively exchanged amino acids (for definition see "Materials and Methods") are *shadowed gray*, identical amino acids are indicated with an *asterisk*. Gaps indicated by points were introduced for optimal alignment of all three Ig-like domains. The *inset* shows the schematic domain organization of the aligned protein.

rat CEA-like gene rnCGM1. Furthermore, comparison of the derived amino acid sequence of the rat protein with the sequence of the human CEA-related cross-reacting antigen biliary glycoprotein 1 (Hinoda *et al.*, 1988; Barnett *et al.*, 1989) demonstrated that it is organized in a number of repeated domains. Surprisingly, the domain that is encoded by the N-terminal or IgV-like domain exon in the human CEA-like genes, which comprises the last third of the leader sequence and the complete N-terminal domain, is repeated five times in the putative rat protein (Figs. 1 and 3), whereas this domain is present only in one copy in the human CEA-like proteins so far analyzed (Thompson *et al.*, 1989a). These five repeated domains show a similarity to each other which lies between 61 and 89% at the nucleotide level, and between 41 and 76% at the amino acid level. The highest values are found between domains N_2 and N_3 (Table I). Furthermore, the last 93 amino acids encoded by the cDNA correspond to a half-repeat type A, which is related to the C-like domain of the immunoglobulins (Fig. 3). Analysis of the deduced amino acid sequence for putative *N*-glycosylation consensus sequences reveals 16 sites, all contained in the IgV-like domains. This rather large number of potential glycosylation sites suggests that this rat CEA-like protein is highly glycosylated, but probably less than CEA, which has a similarly sized protein core and 28 putative glycosylation sites (Oikawa *et al.*, 1987a).

Determination of the Exon Organization and Transcriptional Starts of the rnCGM1 Gene—In order to prove the unexpected domain structure of the rat CEA-like protein, the corresponding genomic clone λrnCGM1-1 was analyzed. The localization of the first six exons present in clone λrnCGM1-1 was performed by restriction mapping and sequence determination of the exons and adjacent intron regions (Fig. 1), which are flanked by canonical splice donor and acceptor sequences (Table II). The first exon encodes the 5'-untranslated region and about two-thirds of the leader peptide. The following five exons each code for the last third of the leader

λrnCGM1-1 which contains the first exon encoding two-thirds of the leader and the 5'-noncoding region. The two alternative start points of transcription as determined by primer extension are marked by *arrows*. The *arrowhead* indicates the exon/intron border.

TABLE I

Sequence comparison of IgV-like domain exons of CEA-related genes of the rat

Amino acid and nucleotide sequence comparison of the IgV-like domain exons of rnCGM1 (N_1-N_5) and the analyzed exons of the genomic clones λrnCGM2-1, rnCGM3-1, rnCGM4/5-1 exon A (rnCGM4A) and rnCGM4/5-1 exon B (rnCGM4B) (Kodelja *et al.*, 1989) are shown. The percent similarity was calculated after optimal alignment.

	Amino acid level								
	N1	N2	N3	N4	N5	rnCGM2	rnCGM3	rnCGM4A	rnCGM4B
N1		51	51	41	48	45	76	73	49
N2	66		76	46	48	44	42	52	65
N3	64	89		46	51	44	42	49	63
N4	62	62	61		44	33	39	41	53
N5	64	64	67	61		42	45	49	54
rnCGM2	60	61	60	54	60		39	47	46
rnCGM3	84	61	61	57	60	56		64	45
rnCGM4A	80	66	64	61	63	62	72		51
rnCGM4B	65	76	75	65	67	60	61	65	
	Nucleotide level								

TABLE II

Nucleotide sequences of exon/intron borders of the rnCGM1gene

Capital letters represent coding sequences, lower case letters represent intron sequences. Numbers refer to corresponding positions in the cDNA. Abbreviations for the exons are as in the legend to Fig. 1.

Junction	Exon	Donor	Intron	Acceptor	Exon
L-N1	...CTG CTC ACA Leu Leu Thr	·64 G gtaagggtgctt A	.....	ttttctctcttcccttctag	·65 CC TCC CTC TTA... la Ser Leu Leu
N1-N2	...CAT GTG TAC His Val Tyr	·424 T gtaagtaattct P	.....	gatctctgtcttccttctag	·425 TC TCC ACT TTG... he Ser Thr Leu
N2-N3	...CAG GTG GAC Gln Val Asp	·784 A gtaagtagttct T	.....	atctctctcttcttttctag	·785 CC TGT TTT ATG... hr Cys Phe Met
N3-N4	...CAG GTG AAC Gln Val Asn	·1144 A gtaagtgaatct T	.....	ttttctcctttcccttccag	·1145 CC TCC TCG TGC... hr Ser Ser Cys
N4-N5	...AGC GTG CAC Ser Val His	·1498 A gtaagtgactct T	.....	ttttctatctgccctttag	·1499 CC TGT GTT CAC... hr Cys Val His
N5-(R_A)	...CAC ATA TAC His Ile Tyr	·1846 A gtaagtaattct L(ys)	.....		

or a leader-like sequence, respectively, followed by the N-terminal or IgV-like domain (Fig. 1, Table II). The exons encoding the IgC-like domain and the 3′-untranslated region are not included in λrnCGM1-1. No additional IgV- or IgC-like domain exons could be identified by Southern hybridization on λrnCGM1 (data not shown). In addition, we sequenced the putative promoter region covering 703-bp upstream of the start codon (Fig. 2*B*). The transcription initiation site was determined by primer extension. Two extension products were found, which correspond to start sites at position −192 and −189 nucleotides upstream of the initiation codon. This is based on the assumption that no intron is present in the 5′-untranslated region between the determined transcriptional start and the 5′-end of the rnCGM1a cDNA. Indeed, no recognizable intron acceptor consensus sequence can be found in this region (Fig. 2*B*). In the vicinity upstream of the transcriptional start sites, no classical consensus sequences for TATA- and CAAT-boxes could be identified (Cordon *et al.*, 1980).

Northern Blot Analyses—For the characterization of the mRNA(s) transcribed from the gene rnCGM1, day 21 rat placental total RNA was hybridized in Northern experiments with the complete *Eco*RI cDNA insert of clone λrnCGM1b. Three mRNA species with a length of 3.9, 3.2, and 2.5 kb could be detected under low stringency hybridization conditions (Fig. 4*A*). After washing at higher stringency (0.1 × SSPE, 65 °C) the smallest band disappeared almost completely but the others remained (data not shown). To test the possibility that cross-hybridization of the probe with the 3.9- and 3.2-kb mRNA species could be due to expression of two closely related genes, a probe covering the 3′-noncoding region was used. This region is known to be completely different among certain CEA-related mRNA species in man (Zimmermann *et al.*, 1988). With this probe, only the 3.9- and 3.2-kb mRNA from day 21 rat placenta poly(A) RNA were found to hybridize even at low stringency conditions (Fig. 4*B*). This result suggests that both mRNA species correspond to the same gene. The size difference could be caused by either alternative polyadenylation as has been found for human CEA mRNA (Oikawa *et al.*, 1988), or differential splicing which has been described for a number of CEA-like genes in man (Hinoda *et al.*, 1988; Barnett *et al.*, 1989; Khan and Hammarström, 1989; Khan *et al.*, 1989; Watanabe and Chou, 1988; Streydio *et al.*, 1988; Zimmermann *et al.*, 1989). In order to identify tissues which express the gene rnCGM1, we screened a number of fetal and adult tissues. The knowledge of the expression pattern should eventually help to assign this member of a rapidly diverging gene family to its human counter-

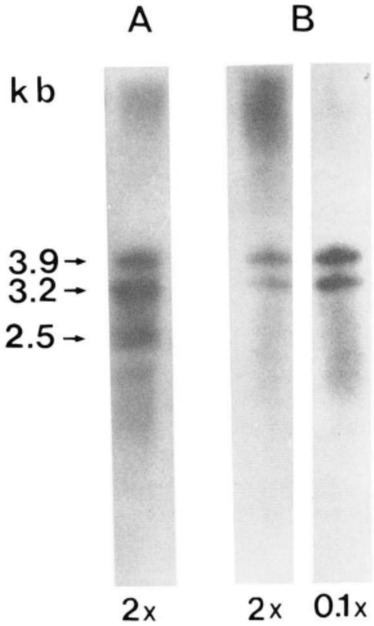


FIG. 4. **Identification of rnCGM1 mRNA species.** *A*, 10 μg of total RNA or, *B*, 2 μg of poly(A) RNA from a day-21 rat placenta were size-fractionated on a 1% agarose gel containing (*A*) 2.2 M formaldehyde or (*B*) in the presence of methylmercury hydroxide, respectively, blotted onto a charged nylon membrane, and hybridized with the ^{32}P-labeled cDNA fragments. As probes either the complete insert of λrnCGM1b (*A*) or the 1073-bp *Eco*RI/*Sst*I fragment of λrnCGM1b covering the 3′-noncoding region (*B*) were used. The blots were washed under low (2 × SSPE, 60 °C) or high stringency conditions (0.1 × SSPE, 65 °C), respectively. Autoradiography was carried out overnight (*B, lane 1*), for three days (*A*) or for eight days (*B, lane 2*).

part. Except in placenta, neither the 3.9- and 3.2-kb transcripts of rnCGM1 nor the 2.5-kb mRNA species could be detected in the analyzed fetal (liver, intestine, brain) and adult tissues (small intestine, stomach, colon tumor, kidney, lung (data not shown)). Furthermore, Northern blot experiments under low stringency conditions with RNA from placentae of day 14 to 21 of gestation using the complete insert of clone λrnCGM1b as a probe revealed a constant steady state level, both absolute and relative, of all three mRNA species described above. This was verified by hybridization with a rat β-actin probe (data not shown).

DISCUSSION

In this paper we have characterized the primary structure of a CEA-related protein in rat (rnCGM1) by cDNA analysis. It shows a strikingly different domain organization to the homologous human CEA family members. In contrast to all human CEA-like proteins, which are composed of only one IgV-like or N-terminal domain and a varying number (2, 3, 6) of IgC-like or half-repeat domains (Thompson *et al.*, 1989a), the rnCGM1 protein contains five IgV-like domains (N_1–N_5), separated by truncated leader-like sequences (L_2'–L_5') of variable length, and only one IgC-like or half-repeat domain (Figs. 1 and 3). The latter domain can be clearly classified as an A-type half-repeat domain (R_A) by comparison with human R_A and R_B sequences (Fig. 3). This domain structure is reminiscent of the domain organization of the rabbit poly Ig receptor, an integral membrane protein with four IgV-like domains, which have only a marginal sequence similarity to each other, one IgC-like domain, but no recognizable internal leader "remnants" (Mostov *et al.*, 1984; Williams and Barclay, 1988). However, sequence comparison revealed, that rnCGM1 is no more related to the poly Ig receptor than to other members of the Ig superfamily. The IgV-like domains of the rnCGM1 protein all contain a pair of oppositely charged amino acids (Arg, Asp (Fig. 3)) at conserved positions, characteristic for the N-terminal domains of all CEA-like proteins analyzed so far. It is supposed that they form a salt bridge which is thought to stabilize the β-sheet structure of the Ig-fold in the absence of a disulfide bond (Williams, 1987; Williams and Barclay, 1988). Furthermore, the domains N_1–N_5 show the same subdivision in conserved and hypervariable regions as found in other rat and human N-terminal domains (Fig. 3; Kodelja *et al.*, 1989; Thompson *et al.*, 1989a). The unexpected organization of N-terminal domains in the rnCGM1 protein has been confirmed by determination of the exon structure of the corresponding rat gene (Fig. 1). The exon/intron borders are identical to the ones in the human CEA-like genes and correspond well to the domain borders (Thompson *et al.*, 1987, 1989b; Oikawa *et al.*, 1987b; Barnett *et al.*, 1989). Additional proteins with a similar structure can be predicted to exist in rat as inferred from the structure of another partial rat genomic clone (λrnCGM4/5) which contains two adjacent N-terminal domain exons separated by a 2924-bp intron (Kodelja *et al.*, 1989).[4] No further exons could be found in this intron by determination of its complete nucleotide sequence. Comparing the nucleotide and amino acid sequences of the N-terminal domain exons of λrnCGM4/5 with N_1–N_5 of rnCGM1, the highest similarity is found between λrnCGM4/5-1 exon A and N_1 and λrnCGM4/5 exon B and N_2 (Table I). This suggests that these two exons encode the first two N-terminal domains of another rat CEA-related protein. A similarly high conservation of flanking intron sequences support this (data not shown). Furthermore, we have recently isolated two partial clones from a murine placental cDNA library which code for different proteins with at least two adjacent N-terminal domains each.[5] As a consequence of this unexpected organization, the N-terminal domain exons cannot be used to count the number of CEA-related genes in rodents as has been done in man. Therefore, presently it is not clear how many genes are represented by the five rat N-terminal domain exons, whose sequences have been published recently (Kodelja *et al.*, 1989).

Taken together these rat genes with multiple N-terminal domain exons have probably evolved by multiplication of N-terminal domain exons rather than by multiplication of whole N-terminal and half-repeat containing genes. The oligomerization has probably occurred stepwise, whereby the N_2 and N_3 exons have arisen rather late in evolution. This is inferred from the relatively high similarity between the sequences of N_2 and N_3 (Table I) and from the low rate of synonymous substitutions observed when the nucleotide sequences of the N_2 and N_3 domains are compared (K_s = 0.13 ± 0.04). In contrast, all other comparisons with different N-terminal domain combinations lead to much higher K_s values (K_s = 0.82 ± 0.18 to 1.18 ± 0.31). Synonymous substitutions do not result in amino acid exchanges. Therefore, no selective pressure is exerted at such silent sites and their mutation frequencies are proportional to the time of sequence divergence.

The fact that at least some rodent and human CEA-like proteins show a different domain organization may have important functional implications. As discussed below, the rnCGM1 protein probably represents a rat counterpart to a human PSG. Assuming a similar function for the PSG-like proteins in rat and man, the number of IgV-like or IgC-like

[4] K. Lucas and W. Zimmermann, unpublished results.

[5] F. Rudert, W. Zimmermann, and J. A. Thompson, unpublished results.

domains does not appear to be of functional importance, but rather would appear to determine the shape or length of the protein. This could be achieved by a serial arrangement of IgV- or IgC-like domains which exhibit only a low sequence similarity. The first IgV-like domain possibly conveys the specific function of the various CEA-like proteins because such a domain type is present at the amino-terminal end of all CEA-like molecules analyzed so far. This hypothesis is strengthened by the observation that the members of the CEA gene family show a rapid interspecies sequence divergence during evolution (Rudert *et al.*, 1989). This is highlighted by the fact that for all genes the amino acid sequences are less conserved than the nucleotide sequences. Furthermore, the intron and 3′- and 5′-noncoding nucleotide sequences of the CEA gene family members show, in general, similar conservation to the coding sequences (Rudert *et al.*, 1989, Zimmermann *et al.*, 1989, Thompson *et al.*, 1989b). All these facts indicate low selective pressure for the overall amino acid sequence. Rather, they suggest that only a few, critical amino acids (see Fig. 3) must remain conserved in order to guarantee a fixed secondary (β-sheet) and tertiary structure (Ig-like fold). A similarity for corresponding domains ranging from 33 to 52% at the amino acid level is found with members of both human subgroups. For these reasons, it is not possible to assign by sequence comparison, the rnCGM1 protein to its human counterpart or even to the CEA- or PSG-subgroups. A partial cDNA clone from a murine adult colon library (Beauchemin *et al.*, 1989) which covers one IgV and two IgC domains, shows a higher degree of similarity to the human CEA/PSG gene family (47–67%). Despite this, it also cannot be assigned to one or the other human subgroup. Recently, however, it has been reported by Lin and Guidotti (1989) that a rat liver plasma membrane eco-ATPase is the counterpart to the human biliary glycoprotein 1, a member of the CEA family (Hinoda *et al.*, 1988; Barnett *et al.*, 1989). Despite a low overall sequence similarity (56% at the amino acid level) this rat CEA family member shows the same tissue distribution (Lin, 1989; Svenberg, 1976) and in contrast to rnCGM1, the same domain organization as its human counterpart biliary glycoprotein 1.

The tissue distribution of rnCGM1 mRNA is reminiscent of the one found for the mRNAs of the human PSG genes, most of which have been demonstrated to be strongly expressed in the placenta (Watanabe and Chou, 1988; Chan *et al.*, 1988; Chou *et al.*, 1989; Zimmermann *et al.*, 1989), although minor expression has also been described for other tissues (Chan *et al.*, 1988). Furthermore, the lack of any hydrophobic domain suggests that the rnCGM1 protein is directly secreted. This is also assumed to be the case with the human PSGs, which have, in contrast to the CEA-subgroup members, in most cases only very short, not particularly hydrophobic C-terminal domains (Rooney *et al.*, 1988, Streydio *et al.*, 1988; Chan *et al.*, 1988; Oikawa *et al.*, 1988; Khan and Hammarström, 1989; Khan *et al.*, 1989). Indeed, human PSGs accumulate in large amounts during pregnancy in the maternal serum (Lin *et al.*, 1974) and PSGs expressed from transfected cDNAs are found in the media (Khan and Hammarström, 1989; Khan *et al.*, 1989; Zimmermann *et al.*, 1989). Together these findings imply that most if not all PSGs are secreted. Recently, the major PSG from rat placenta has been identified to be an ~120-kDa glycoprotein synthesized in basal zone tissue which appears to be secreted (Ogilvie *et al.*, 1989). This size would be in good agreement with the calculated M_r of the mature rnCGM1 protein (M_r = 135,000), assuming a M_r for the carbohydrate residues/*N*-glycosylation site to be similar to CEA, which has been estimated to be ~4,000 (Neumaier *et al.*, 1988). The former calculation implies that only the first leader of the rnCGM1 protein serves as a signal for the signal peptidase and is cotranslationally removed. The internal leader fragments, on the other hand, are probably too short and their amino acid sequences too poorly conserved (Fig. 3) to be recognized by the signal peptidase (Heinje, 1986). If this were the case, the cysteine residues in each of the internal leader-like segments could form intermolecular disulfide bridges leading to the formation of homo- or heterodimers as found for other members of the Ig-superfamily (Williams and Barclay, 1988). Analysis of the rnCGM1 protein through transfection experiments in eucaryotic cells should clarify its processing and modification. In conclusion, rnCGM1 seems to be homologous to members of the human PSG subgroup. Final assignment of the rnCGM1 cDNA to a rat PSG, however, will only be possible through partial amino acid sequencing of the latter. Further cDNA analyses have to be performed to resolve the question whether the domain organization of rnCGM1 is representative for the rat PSG subfamily or not.

Acknowledgments—We wish to thank S. von Kleist for laboratory use and M. Weiss for her excellent technical assistance. The gift of the rat placental cDNA library by M. L. Duckworth is gratefully acknowledged. The oligonucleotides were synthesized by G. Igloi on an automated DNA synthesizer, which was a gift from the Fonds der Chemischen Industrie to H. Kössel.

REFERENCES

Alwine, J., Kamp, D. J., and Stark, G. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5350–5354

Aviv, H., and Leder, P. (1972) *Proc. Natl. Acad. Sci. USA* **69,** 1408–1412

Barnett, T. R., Kretschmer, A., Austen, D. A., Goebel, S. J., Hart, J. T., Elting, J. J., and Kamarck, M. E. (1989) *J. Cell Biol.* **108,** 267–276

Beauchemin, N., Turbide, C., Afar, D., Bell, J., Raymond, M., Stanners, C.-P., and Fuks, A. (1989) *Cancer Res.* **49,** 2017–2021

Benchimol, S., Fuks, A., Jothy, S., Beauchemin, N., Shirota, K., and Stanners, C. P. (1989) *Cell* **57,** 327–334

Chan, W.-Y., Borjigin, J., Zheng, Q.-X., and Shupert, W. L. (1988) *DNA* **7,** 545–555

Chan, W. Y., Tease, L. A., Borjigin, J., Chan, P. K., Rennert, O. M., Srinivasan, B., Shupert, W. L., and Cook, R. G. (1988) *Human Reprod.* **3,** 677–685

Chou, J. Y., Sartwell, A. D., Wan, Y.-J. Y., and Watanabe, S. (1989) *Mol. Endocrinol.* **3,** 89–96

Cordon, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C., and Chambon, P. (1980) *Science* **209,** 1406–1414

Druckrey, H. (1971) *Arzneim. Forsch.* **21,** 1274–1278

Feinberg, A. P., and Vogelstein, B. (1983) *Anal. Biochem.* **132,** 6–13

Feng, D. F., Johnson, M. S., and Doolittle, R. F. (1985) *J. Mol. Evol.* **21,** 112–125

Fiddes, J. C., and Goodman, H. M. (1979) *Nature* **281,** 351–356

von Heijne, G. (1986) *Nucleic Acids Res.* **14,** 4683–4690

Hinoda, Y., Neumaier, M., Hefta, S. A., Drzenick, Z., Wagener, C., Shively, L., Hefta, L. J., Shively, J. E., and Paxton, R. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 6959–6963. Correction: *Proc. Natl. Acad. Sci. USA* (1989) **86,** 1668

Khan, W. N., and Hammarström, S. (1989) *Biochem. Biophys. Res. Commun.* **161,** 525–535

Khan, W. N., Osterman, A., and Hammarström, S. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 3332–3336

Kodelja, V., Lucas, K., Barnert, S., von Kleist, S., Thompson, J. A., and Zimmermann, W. (1989) *J. Biol. Chem.* **264,** 6906–6912

Li, W.-H., Wu, C.-I., and Luo, G.-C. (1985) *Mol. Biol. Evol.* **2,** 150–174

Lin, S.-H. (1989) *J. Biol. Chem.* **264,** 14403–14407

Lin, S.-H., and Guidotto, G. (1989) *J. Biol. Chem.* **264,** 14408–14414

Lin, T.-M., Halpert, S. P., and Spellacy, W. N. (1974) *J. Clin. Invest.* **54,** 576–582

Messing, J. (1983) *Methods Enzymol.* **101,** 20–78

Mostov, K. E., Friedlander, M., and Blobel, G. (1984) *Nature* **308,** 37–43

Neumaier, M., Zimmermann, W., Shively, L., Hinoda, Y., Riggs, A.

D., and Shively, J. E. (1988) *J. Biol. Chem.* **263,** 3202–3207

Ogilvie, S., Kvello-Stenstrom, A. G., Hammond, G., Buhi, W. C., Larkin, L. H., and Shiverick, K. T. (1989) *Endocrinology* **125,** 287–294

Oikawa, S., Nakazato, H., and Kosaki, G. (1987a) *Biochem. Biophys. Res. Commun.* **142,** 511–518

Oikawa, S., Kosaki, G., and Nakazato, H. (1987b) *Biochem. Biophys. Res. Commun.* **146,** 464–469

Oikawa, S., Inuzuka, C., Kosaki, G., and Nakazato, H. (1988) *Biochem. Biophys. Res. Commun.* **156,** 68–77

Rooney, B. C., Horne, C. H. W., and Hardman, N. (1988) *Gene (Amst.)* **71,** 439–449

Rudert, F., Zimmermann, W., and Thompson, J. (1989) *J. Mol. Evol.* **29,** 126–134

Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467

Shiveley, J. E., and Beatty, J. D. (1985) *CRC Crit. Rev. Oncol./Hematol.* **2,** 355–399

Streydio, C., Lacka, K., Swillens, S., and Vassart, G. (1988) *Biochem. Biophys. Res. Commun.* **154,** 130–137

Svenberg, T. (1976) *Int. J. Cancer* **17,** 588–596

Tatarinov, Y. S. (1978) *Gynecol. Obstet. Invest.* **9,** 65–97

Thompson, J., Barnert, S., Berling, B., von Kleist, S., Kodelja, V., Lucas, K., Mauch, E.-M., Rudert, F., Schrewe, H., Weiss, M., and Zimmermann, W. (1989a) in *The Carcinoembryonic Antigen Gene Family* (Yachi, A. and Shively, J. eds) pp. 65–74, Elsevier Science Publishers BV, Amsterdam

Thompson, J. A., Mauch, E.-M., Chen, F.-S., Hinoda, Y., Schrewe, H., Berling, B., Barnert, B., von Kleist, S., Shively, J. E., and Zimmermann, W. (1989b) *Biochem. Biophys. Res. Commun.* **158,** 996–1004

Thompson, J. A., Pande, H., Paxton, R. J., Shively, L., Padma, A., Simmer, R. L., Todd, C. T., Riggs, A. D., and Shively, J. E. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 2965–2969

Watanabe, S., and Chou, J. Y. (1988) *J. Biol. Chem.* **263,** 2049–2054

Williams, A. F. (1987) *Immunol. Today* **8,** 298–303

Williams, A. F., and Barclay, A. N. (1988) *Annu. Rev. Immunol.* **6,** 381–405

Young, R. A., and Davis, R. W. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1194–1198

Zimmermann, W., Weber, B., Ortlieb, B., Rudert, F., Schempp, W., Fiebig, H.-H., Shively, J. E., von Kleist, S., and Thompson, J. A. (1988) *Cancer Res.* **48,** 2550–2554

Zimmermann, W., Weiss, M., and Thompson, J. A. (1989) *Biochem. Biophys. Res. Commun.* **163,** 1197–1209