Molecular and Cellular Biology, June 1990, p. 2738–2748
0270-7306/90/062738-11$02.00/0

Vol. 10, No. 6

Cloning of the Complete Gene for Carcinoembryonic Antigen: Analysis of Its Promoter Indicates a Region Conveying Cell Type-Specific Expression

HEINRICH SCHREWE,[1] JOHN THOMPSON,[1]* MARION BONA,[1] LAURA J. F. HEFTA,[2] AIKO MARUYA,[2] MARTIN HASSAUER,[1] JOHN E. SHIVELY,[2] SABINE VON KLEIST,[1] AND WOLFGANG ZIMMERMANN[1]

Institut für Immunbiologie der Universität, Stefan-Meier-Strasse 8, D-7800 Freiburg, Federal Republic of Germany,[1] *and Beckman Research Institute of the City of Hope, Duarte, California 91010-0269*[2]

Received 21 December 1989/Accepted 1 March 1990

Carcinoembryonic antigen (CEA) is a widely used tumor marker, especially in the surveillance of colonic cancer patients. Although CEA is also present in some normal tissues, it is apparently expressed at higher levels in tumorous tissues than in corresponding normal tissues. As a first step toward analyzing the regulation of expression of CEA at the transcriptional level, we have isolated and characterized a cosmid clone (cosCEA1), which contains the entire coding region of the CEA gene. A close correlation exists between the exon and deduced immunoglobulin-like domain borders. We have determined a cluster of transcriptional starts for CEA and the closely related nonspecific cross-reacting antigen (NCA) gene and have sequenced their putative promoters. Regions of sequence homology are found as far as approximately 500 nucleotides upstream from the translational starts of these genes, but farther upstream they diverge completely. In both cases we were unable to find classic TATA or CAAT boxes at their expected positions. To characterize the CEA and NCA promoters, we carried out transient transfection assays with promoter-indicator gene constructs in the CEA-producing adenocarcinoma cell line SW403, as well as in nonproducing HeLa cells. A CEA gene promoter construct, containing approximately 400 nucleotides upstream from the translational start, showed nine times higher activity in the SW403 than in the HeLa cell line. This indicates that *cis*-acting sequences which convey cell type-specific expression of the CEA gene are contained within this region.

The carcinoembryonic antigen (CEA) was originally described as a glycoprotein molecule with an oncofetal expression pattern (13). Recent experiments indicate that CEA may function as a cell adhesion molecule, which could play an important role during embryogenesis and possibly also during tumor development (5). Despite its presence in some normal tissues, its concentration in serum is a clinically useful parameter, especially in the postoperative monitoring of colonic tumor patients (52). CEA is a member of a family of closely related molecules, whose genes reveal a high degree of sequence similarity (reviewed in reference 49). The CEA family shows structural resemblance to, and can be placed within, the immunoglobulin superfamily (4, 32, 37). The CEA family members are made up of one N-terminal immunoglobulin variable-like, and a varying number of immunoglobulin constant-like domains, according to the classification of Williams (57). The CEA gene family can be divided into two main subgroups based on sequence comparisons. One subgroup contains the CEA gene itself (2, 4, 36, 59) and those encoding the classic CEA immunocross-reacting antigens, such as the nonspecific cross-reacting antigens (NCA) (2, 31, 35, 47, 51) and biliary glycoprotein (3, 17). These molecules are membrane bound, either as integral proteins such as biliary glycoprotein or after posttranslational modification through a covalently linked glycosyl phosphatidylinositol moiety as shown for CEA (16, 19, 45) and an NCA (15, 21). The second subgroup contains the genes encoding the pregnancy-specific glycoproteins (PSG), which were recently shown to have homology to the CEA subgroup members (56). These proteins are expressed at high levels in the placenta (26) and at lower levels in trophoblastic tumors (46).

Despite the high sequence similarity of CEA and PSG, they show differential expression patterns. PSG mRNAs are found in the placenta, where no apparent expression of any of the CEA gene subgroup mRNAs can be determined (48, 61). Colonic tumors, on the other hand, contain transcripts for CEA and NCA but not PSGs (48, 60, 61). In addition, differential expression of CEA and NCA mRNAs can be found in various other tumors (7, 60). There is a notable difference in the amounts of CEA protein found in normal colonic mucosa compared with colonic tumors (8). Some evidence has recently been presented which would suggest that a rapid loss of CEA takes place in normal colonic mucosa, but not in the corresponding tumors, and that the rates of synthesis are comparable in both cases, which would explain these differences (23). However, other results suggest an increased transcriptional activity in colonic tumors compared with the normal mucosa (7), which would also suggest regulation of expression at the transcriptional level. Indeed, differential methylation patterns could be shown to correlate with the rate of CEA expression in various cell lines (54). Recently, other factors have been reported to exist which also play a role in the regulation of transcription and translation of CEA-related genes (53).

As a basis for more detailed studies on the transcriptional regulation of tissue-specific and developmental expression of CEA, as well as its possible role during embryo and tumor development, it is necessary to characterize the CEA gene. In this paper, we report the isolation of a cosmid clone which contains the complete transcriptional unit of the CEA gene, including its promoter region. We have carried out restriction endonuclease mapping and have determined the exon

---

* Corresponding author.

structure of this gene. Sequence and functional analyses of the putative promoter regions of the CEA gene, together with an NCA gene whose isolation we have described elsewhere (51), have been carried out to locate their promoters. Furthermore, we have identified a region which apparently conveys tissue-specific expression of the CEA gene.

MATERIALS AND METHODS

Chemicals and oligonucleotides. Enzymes were purchased from Boehringer Mannheim Biochemicals, Bethesda Research Laboratories, Inc., and Pharmacia/LKB. Radiochemicals, RNasin, and Taq polymerase were from Amersham Corp. and Du Pont, NEN Research Products. All other chemicals were of analytical grade. Oligonucleotides were synthesized by the phosphoramidite method on an Applied Biosystems 308A DNA synthesizer and purified by following the manufacturer's protocol, as described previously (20).

Bacterial strains and cloning vectors. All work with recombinant DNA was carried out in accordance with the German and U.S. safety regulations. A human cosmid genomic library inserted in the vector pHC79-2cos/tk (27) was obtained as packaged cosmids from W. Lindenmaier, Gesellschaft für Biotechnologische Forschung, Brunswick, Federal Republic of Germany, and transduced into *Escherichia coli* ED8767 by combining 0.1 ml of 50 mM Tris hydrochloride (pH 7.5), 10 mM $MgSO_4$, 0.1 ml of ED8767 in LB medium (optical density at 600 nm = 2) and 0.1 ml of packaged cosmids. After incubation for 30 min at 30°C the suspension was diluted with 2 ml of LB medium, incubated for 60 min at 37°C with shaking, and plated out. For subcloning of the insert DNA fragments, either the Bluescript phagemid (Stratagene Inc.) and pUC18/19 were used in *E. coli* JM109 and RR1ΔM15, or M13mp18/19 bacteriophage DNA was transfected and amplified in *E. coli* JM107.

Screening of the human genomic cosmid library. Recombinant clones were grown on nitrocellulose (Schleicher & Schuell, Inc.) in the presence of ampicillin (50 μg/ml) and amplified after the addition of chloramphenicol (150 μg/ml). Colony hybridization with a CEA-specific, ^{32}P-labeled, 408-base-pair (bp) *Rsa*I-*Pst*I restriction endonuclease fragment from the 3′ untranslated region of a CEA cDNA clone (pCEA3 [60]) was performed in the presence of 50% formamide, 6× SSPE (1× SSPE is 0.18 M NaCl, 10 mM NaH_2PO_4 [pH 7.4], and 1 mM EDTA), 1× Denhardt solution (0.02% each Ficoll, polyvinylpyrrolidone, and bovine serum albumin), 0.05% sodium phosphate, and 100 μg of sheared, denatured salmon sperm DNA per ml at 42°C. The final wash was in 2× SSPE–0.5% sodium dodecyl sulfate at 65°C.

Positive clones were purified and, for identification of a full-length CEA gene clone, were hybridized with a ^{32}P-labeled synthetic oligonucleotide (CEAL, 5′-AGCCTCTGC CAGGGGATGCACCAT-3′) corresponding to nucleotide positions 1798 to 1821 of the leader region of the CEA gene (see Fig. 2). The hybridization took place in 50% formamide–1 M NaCl–1% sodium dodecyl sulfate–100 μg of sheared, denatured salmon sperm DNA per ml–10% dextran sulfate at 30°C, and washing was carried out in 2× SSPE–1% sodium dodecyl sulfate at 10°C below the calculated melting temperature (24).

Restriction endonuclease mapping. The cosmid clone was mapped by digestion with the restriction endonuclease *Aat*II, which has three recognition sites in the vector. A 32.6-kilobase-pair (kb) *Aat*II fragment contains the whole of the insert and flanking vector sequences (0.1 and 1 kb). This fragment was then partially digested with different restriction endonucleases (*Bam*HI, *Sal*I, *Sma*I, *Sst*I, and *Xba*I), separated by gel electrophoresis in 0.5% agarose overnight at 1 V/cm (4°C), and transferred to GeneScreen Plus membrane (Du Pont, NEN) prior to hybridization with an oligonucleotide probe, whose complementary sequence is located in the 0.1-kb flanking sequence (cos1, 5′-TAGGCGTAT CACGAGGCCCTTTCG-3′). After stripping, the blot was rehybridized with an oligonucleotide (cos2, 5′-GGCGATGC TGTCGGAATGGACG-3′) whose corresponding sequence is located in the 1-kb flanking sequence. After autoradiography, the linear arrangement of restriction sites in the partial digests was determined from the fixed *Aat*II sites. Uncertainties arising from this method could be resolved by isolation of selected DNA fragments, followed by digestion with a second restriction endonuclease and by agarose gel electrophoresis separation.

The N-terminal domain and 5′ untranslated regions were located by Southern blot analysis with the 547-bp *Pst*I fragment from pNCA1 (20) and the CEAL oligonucleotide (see above) from the leader region of the CEA cDNA, respectively, as probes. The exon/intron boundary-containing fragments were located on Southern blots by using oligonucleotide probes homologous to regions at the beginning and end of each A and B immunoglobulin constant-like domains. The cosCEA1 DNA was digested with various restriction endonucleases and transferred to Zetaprobe membrane (Bio-Rad Laboratories) as suggested by the manufacturer. The membranes were hybridized with the various oligonucleotides in 5× SSC (1× SSC is 0.015 M sodium citrate plus 0.15 M NaCl)–0.5% nonfat dry milk–0.01% sodium dodecyl sulfate. The hybridization and wash temperatures were calculated by the method of Wallace and Miyada (55). The CEA 3′ untranslated region was located by using an approximately 450-bp *Rsa*I-*Eco*RI cDNA fragment downstream of the first Alu sequence (36).

Determination of DNA sequences. Subcloned restriction endonuclease DNA fragments in phage M13mp18/19, pUC18/19, and Bluescript phagemid vectors were sequenced as single- or double-stranded templates by the dideoxy-chain termination method (40), with universal or internal oligonucleotide primers and a kit from United States Biochemical Corp. For comparison of the nucleotide and deduced amino acid sequences, the computer program Align (M. Trippel and R. Friedrich, unpublished data) was used. The DNA sequence data have been forwarded to GenBank and have received the accession numbers M31966 to M31975.

Determination of the transcriptional starts of the CEA and NCA genes. (i) Primer extension experiments. Gene-specific oligonucleotides complementary to CEA and NCA mRNA species were used to prime extension reactions. We synthesized oligonucleotides from the leader region of the CEA gene (CEAL, see above) and the 5′ untranslated region of the NCA gene (NCA5′; 5′-TCTCTGTCCACCTCTTTG TAGAGGA-3′, corresponding to positions −15 to −39 in Fig. 5). For the annealing reaction, 5 to 10 ng of oligonucleotide labeled with ^{32}P at the 5′ end was mixed with 2.5 μg of poly(A)$^+$ RNA isolated from a colon tumor (59) in 10 μl of 5 mM sodium phosphate buffer (pH 7.0)–5 mM EDTA. After a 5-min denaturation at 90°C, NaCl was added (final concentration, 80 mM), and hybridization took place at 50°C for 1 h. After this time the mixtures were allowed to cool slowly to room temperature. The hybridization mix was adjusted to 17.5 mM Tris hydrochloride (pH 8.3), 4.3 mM $MgCl_2$, 1.75 mM dithiothreitol, 3.5 mM deoxyribonucleoside triphosphates, 1 ng of dactinomycin per μl, 2 U of RNasin per μl, and 0.8 U of reverse transcriptase per μl in a volume of 25 μl

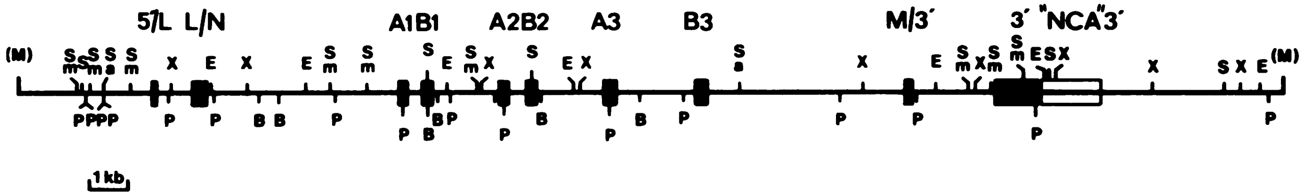

FIG. 1. Restriction map of cosCEA1. The following restriction endonuclease sites have been mapped: *Bam*HI (B), *Eco*RI (E), *Pst*I (P), *Sal*I (Sa), *Sma*I (Sm), *Sst*I (S), *Xba*I (X), and degenerate *Msp*I cloning sites [(M)]. Symbols: ■, exons 5/L through 3'; □, "NCA"3' (this exon has been only provisionally characterized; see text). The exon names shown above the map are explained in the text.

and incubated at 42°C for 1 h. Following phenol extraction and ethanol precipitation, the extension products were denatured and loaded onto a 6% polyacrylamide–7 M urea DNA sequencing gel.

(ii) S1 nuclease analyses. The transcriptional starts of CEA and NCA were independently determined by S1 nuclease mapping. For CEA, a 1.6-kb *Pst*I restriction endonuclease fragment from cosCEA1, containing the first exon of the CEA gene (Fig. 1), and for NCA a 606-bp *Pst*I fragment from clone lambda 39.2 (51), containing the first exon of the NCA gene, were cloned into M13mp18. Gene-specific, single-stranded DNA fragments which were complementary to their mRNAs were synthesized by using the same oligonucleotides as for the primer extension experiments (see above). For this, the oligonucleotides were 5' end labeled with polynucleotide kinase and annealed to the single-stranded M13 subclones by the Perkin Elmer Cetus GeneAmp DNA amplification protocol, followed by extension with Taq polymerase at 72°C for 30 min. The resulting double-stranded products were digested with *Sma*I (for CEA) or *Pst*I (for NCA), and the single-stranded 456nc CEA-specific fragments and the 562nc NCA-specific fragments were separated on a preparative 4% polyacrylamide–7 M urea gel and extracted by electroelution. Hybridization with poly(A)$^+$ RNA from a colonic tumor (see above) and S1 nuclease mapping were performed essentially as described by Maniatis et al. (29). For digestion of single strands, the DNA-RNA complexes were incubated with 1,000 U of S1 nuclease per ml for 30 min at 18°C and analyzed as above.

Construction of promoter-CAT gene hybrids. The 3.4-kb *Sst*I-*Eco*RI fragment of the cosmid clone cosCEA1 (Fig. 1), which contains the first two exons of the CEA gene, was digested with *Nco*I. The 835-bp *Nco*I fragment was isolated and treated with S1 nuclease to generate blunt ends, resulting in an 831-bp DNA fragment located at nucleotide positions −832 to −2 with respect to the translational start of the CEA gene. The vector pBLCAT3 (28), which contains the promoterless bacterial chloramphenicol acetyltransferase (CAT) gene was linearized by using *Bam*HI. The ends were filled in with the Klenow fragment of *E. coli* DNA polymerase I, and the blunt-ended *Nco*I fragment was inserted to generate the constructs pCEA832/2CAT and pCEA2/832CAT (see Fig. 6). A 409-bp *Pst*I-*Bst*XI fragment was removed from clone pCEA832/2CAT, and after generation of blunt ends with T4 DNA polymerase, the rest was religated, resulting in the construct pCEA424/2CAT (see Fig. 6). In parallel, a 420-bp *Hin*dIII-*Bst*XI fragment of plasmid pCEA832/2CAT was removed, and the rest was ligated to a 2.9-kb *Hin*dIII-*Bst*XI fragment from the 5' end of cosCEA1 (Fig. 1), generating the construct pCEA3300/2CAT (see Fig. 6).

The 2.7-kb *Eco*RI fragment from the genomic NCA clone lambda 39.2 (51), which contains the first and second exons of the NCA gene, was subcloned into pUC18. This was digested with *Sst*I, which cuts in the polylinker of pUC18 and internally, generating a fragment containing the 5' untranslated region of the NCA gene, extending from nucleotide positions −48 to −1246 with respect to the translational start (see Fig. 5). This was treated with T4 DNA polymerase to generate blunt ends and inserted into the above-mentioned, blunt-ended pBLCAT3 vector. Clones were constructed containing the inserts in both orientations (pNCA1246/48CAT and pNCA48/1246CAT; see Fig. 6). The construct pNCA1246/48CAT was digested with *Pst*I, the 663-bp fragment from the 5' region of the insert was removed, and the rest was religated, resulting in the construct pNCA583/48CAT (see Fig. 6). A 299-bp *Hin*dIII fragment containing vector sequences and the 5' portion of the insert from clone pNCA583/48CAT was removed. Religation of the rest generated the construct pNCA284/48CAT (see Fig. 6).

Cells and transfection. The colon adenocarcinoma cell line SW403 (25) was grown in RPMI 1640 medium supplemented with 10% fetal calf serum. HeLa cells were cultivated in Dulbecco modified Eagle medium containing 10% fetal calf serum. Cells were plated at a density of 10^6 per 6-cm dish 24 h before transfection. The transfections were carried out by lipofection with 20 μl of Lipofectin (Bethesda Research Laboratories) and 7 μg of pSV2CAT (14) DNA or equimolar amounts of the promoter-CAT gene constructs in 3 ml of cell medium without serum, modified as described by Felgner et al. (12). These parameters were optimized for both cell lines. The cells were incubated for 15 h, and then 3 ml of medium plus 10% fetal calf serum was added; 24 h later the medium was replaced with fresh medium. After a further 24 h of incubation, the cells were harvested in phosphate-buffered saline–1 mM EDTA without bivalent cations, by using a rubber policeman.

CAT assay. Pelleted cells were suspended in 75 μl of 250 mM Tris hydrochloride (pH 7.5), and cell extracts were made by three freeze-thaw cycles. Thereafter, the cells were spun down for 10 min at 13,000 × *g* and 4°C. The supernatants were incubated for 10 min at 65°C, and the CAT activity of the cell extracts was analyzed as described previously (14). The assay mixture contained (in a final volume of 100 μl) 30 μl of cell extract, 20 μl of 8 mM chloramphenicol, 20 μl of 0.5 mM [^{14}C]acetyl coenzyme A (7.4 kBq), and 30 μl of 250 mM Tris hydrochloride (pH 7.5) and was incubated for 1 h at 37°C. After two extractions with 100 μl of ice-cold ethyl acetate to separate the radio labeled acetyl coenzyme A, half of the organic phase containing radiolabeled acetylated chloramphenicol was analyzed by liquid scintillation counting in Aquasol (Du Pont, NEN).

RESULTS

Isolation and characterization of cosCEA1. A CEA-specific cDNA restriction endonuclease fragment from the 3' untranslated region (see Materials and Methods) was used to probe 340,000 cosmid clones from a human genomic library. Three positive clones were isolated whose restriction endonuclease and hybridization patterns with CEA cDNA and oligonucleotide probes from different regions of the gene revealed them to be identical (data not shown). Closer analysis of this clone (cosCEA1) by restriction endonuclease mapping (Fig. 1) and DNA sequence analysis (Fig. 2) revealed that the 31.5-kb insert contains 11 exons, 10 of which represent the complete sequence for the CEA mRNA. The first exon contains the 5′ untranslated region (see below) and part of the leader domain (5′/L exon). The second exon encodes the rest of the leader and the complete N-terminal (immunoglobulin variable-like) domain (L/N exon). The following six exons encode immunoglobulin constant-like half repeats (A and B exons), revealing a strong correlation between the exon and domain borders. A closer analysis of the derived amino acid sequences shows that the immunoglobulin-like domains, which contain a predicted, highly ordered arrangement of β-sheets, are always surrounded by proline-rich regions in CEA. These regions would disrupt ordered structures, thus marking the boundaries between adjacent domains. Such proline-rich regions are located close to the exon borders (Fig. 3). The B3 exon is followed by two further exons: the first contains the hydrophobic membrane domain and part of the 3′ untranslated region (M/3′ exon), and the next contains the rest of the CEA 3′-untranslated region (3′ exon).

Determination of the transcriptional starts of the CEA and an NCA gene. As a basis for closer studies regarding the regulation of transcription of the CEA gene, we have determined its transcriptional start through primer extension (Fig. 4A) and S1 nuclease analyses (Fig. 4B). In both assays, a tight cluster of transcriptional start sites was located 104 to 110 nucleotides upstream from the translational start of the CEA gene (Fig. 4). The primer extension assay indicates transcriptional start positions which are 2 or 3 nucleotides shorter than those determined by the S1 nuclease assay. This difference is probably due to premature termination of the reverse transcriptase in the former assay, presumably because of steric interference by the cap structure, as reported elsewhere for a different gene (18). The location of the transcriptional start site cluster is comparable to that found by Beauchemin et al. (4) for the CEA mRNA. These authors, using only the primer extension analysis determined the transcriptional start to be 102 nucleotides upstream from the translational start.

We have also mapped the transcriptional start for an NCA gene, whose isolation we reported elsewhere (51), by using primer extension and S1 nuclease analyses (Fig. 4). It could be shown that this gene also has a cluster of transcriptional start sites at a similar position to that found for the CEA gene (between nucleotide positions −102 and −112, relative to the translational start).

Sequence comparison of the putative promoter regions with other CEA-related genes. We have sequenced the region around and upstream from the first exon of the NCA genomic clone (Fig. 5). The putative promoter and upstream region of a third gene, CGM1, which was published recently (50), is compared with the NCA sequence in Fig. 5. It is obvious that NCA and CGM1 are homologous over the whole region that was compared. Although no transcripts have so far been identified for the CGM1 gene, this homology indicates that we may expect a similar expression pattern as for NCA, assuming that this gene is active. In comparison, the CEA gene reveals homology in a region up to approximately 500 nucleotides upstream of the initiator codon, but no homology could be determined farther upstream by dot matrix analyses (data not shown).

No canonical TATA or CAAT boxes could be identified at the expected positions (10) upstream of the transcription initiation sites of the CEA or NCA genes. A TATA sequence, which was also described elsewhere (35), can be found approximately 130 nucleotides upstream from the transcriptional start of the NCA gene. However, this distance would appear to be too large to have functional importance, because in most eucaryotic genes TATA sequences are almost always located 25 to 30 bp upstream from the initiation site, except in yeasts, for which distances of up to 120 bp have been described (reviewed in reference 44). Indeed, this putative TATA box is missing in the CEA gene at the corresponding position (Fig. 5), although 2 nucleotides downstream a degenerate TATA sequence can be found (TAGAA).

Analyses of the CEA and NCA gene promoters through transient-transfection assays. To determine the exact locations of the CEA and NCA gene promoters, it was necessary to use a cell line in which the CEA and NCA genes are transcriptionally active. We chose the CEA-producing human adenocarcinoma cell line SW403 (25, 59) and, as a negative control, HeLa cells (59). Northern (RNA) blot analysis revealed transcripts for both CEA and NCA in SW403 cells, but not in HeLa cells (59; data not shown). Various restriction endonuclease fragments, containing up to 3,300 bp of the region upstream from the translational start of the CEA gene, or up to 1,246 bp of the region upstream of the NCA gene translational start, were attached to the bacterial CAT reporter gene. The activities of the various promoter constructs were tested in SW403 and HeLa cells. The relative strengths of the promoter fragments were determined by quantification of the CAT enzyme activity. For comparison, plasmid pSV2CAT, which contains the simian virus 40 early promoter in front of the CAT gene, was transfected into each cell line and CAT activity levels were measured. By using the assumption that the simian virus 40 promoter is equally active in both cell lines, the values obtained for each of the CEA and NCA promoter constructs were expressed as a percentage of the values obtained with pSV2CAT. All of the CEA and NCA constructs were active in their correct orientations, in both cell lines (Fig. 6). Apart from the pNCA1246/48CAT construct, for which the activity in HeLa cells was slightly higher than in SW403 cells, the other NCA constructs showed similar

FIG. 2. Nucleotide sequence of cosCEA1. Intron and exon nucleotides are shown in lowercase and capital letters, respectively. The exon boundaries are marked with arrows over the nucleotide sequences. Restriction endonuclease sites used in subcloning fragments for sequencing are indicated. Polyadenylation sites are underlined. The poly(dA-dC) sequence is heavily underlined. Alu-type repetitive sequences are underlined with dotted lines. The approximate size of nonsequenced regions are shown in brackets and are taken into account for the numbering system (indicated in the margin by an asterisk). The exons were sequenced through the exon/intron boundaries (ca. 100 nucleotides) to confirm identity with the published cDNA sequence of Oikawa et al. (36). In no cases were any substitutions found. A polypurine sequence was noted at the beginning of the gene (1516 to 1588).

```
Sst1                             Pst1                                                                     Pst1
gagctcctca cacggactct gtcagctcct ccctgcagcc tatcggccgc ccacctgagg cttgtcggcc gcccacttga ggcctgtcgg ctgccctctg   100
caggcagctc ctgtccccta caccccctcc ttccccgggc tcagctgaaa gggcgtctcc cagggcagct ccctgtgatc tccaggacag ctcagtctct   200
cacaggctcc gacgccccct atgctgtcac ctcacagccc tgtcattacc attaactcct cagtcccatg aagttcactg agcgcctgtc tcccggttac   300
aggaaaactc tgtgacaggg accacgtctg tcctgctctc tgtggaatcc cagggcccag ccagtgcctg acacggaaca gatgctccat aaatactggt   400
taaatgtgtg ggagatctct aaaaagaaac atatcacctc cgtgtggccc ccagcagtca gagtctgttc catgtggaca caggggcact ggcaccagca   500
                                                                                         Pst1
tgggaggagg ccagcaagtg cccgcggctg ccccaggaat gaggcctcaa cccccagagc ttcagaaggg aggacagagg cctgcaggga atagatcctc   600
           Pst1
cggcctgacc ctgcagccta atcctgagtt cagggtcagc tcacaccacg tcgaccctgg tcagcatccc tagggcagtt ccagacaagg ccggaggtct   700
cctcttgccc tccagggggt gacattgcac acagacatca ctcaggaaac ggattcccct ggacaggaac ctggctttgc taaggaagtg gaggtggagc   800
ctggtttcca tcccttgctc caacagaccc ttctgatctc tcccacatac ctgctctgtt cctttctggg tcctctgagg acctgttctg ccaggggtcc   900
ctgtgcaact ccagactccc tcctggtacc accatgggga aggtggggtg atcacaggac agtcagcctc gcagagacag agaccaccca ggactgtcag  1000
ggagaacatg gacaggccct gagccgcagc tcagccaaca gacacggaga gggagggtcc ccctggagcc ttccccaagg acagcagagc ccagagtcac  1100
ccacctccct ccaccacagt cctctctttc caggacacac aagacacctc cccctccaca tgcaggatct ggggactcct gagacctctg ggcctgggtc  1200
tccatccctg ggtcagtggc ggggttggtg gtactggaga cagagggctg gtccctcccc agccaccacc cagtgagcct tttctagcc  cccagagcca  1300
cctctgtcac cttcctgttg ggcatcatcc caccttccca gagccctgga gagcatgggg agacccggga cctgctgggt ttctctgtca caaaggaaaa  1400
taatccccct ggtgtgacag acccaaggac agaacacagc agaggtcagc actggggaaa gacaggttgt ccacagggga tgggggtcca tccaccttgc  1500
cgaaaagatt tgtctgagga actgaaaata gaagggaaaa aagaggaggg acaaaagagg cagaaatgag aggggagggg acagaggaca cctgaataaa  1600
                                                                  ┌─── 5'UT
gaccacaccc atgacccacg tgatgctgag aagtactcct gccctaggaa gagactcaGG GCAGAGGGAG GAAGGACAGC AGACCAGACA GTCACAGCAG  1700
                                                                         ┌─── Leader
CCTTGACAAA ACGTTCCTGG AACTCAAGCT CTTCTCCACA GAGGAGGACA GAGCAGACAG CAGAGACCAT GGAGTCTCCC TCGGCCCCTC CCCACAGATG  1800
                                                                  Me tGluSerPro SerAlaProP roHisArgTr
                                 ┌──Intron 1
GTGCATCCCC TGGCAGAGGC TCCTGCTCAC AGgtgaaggg aggacaaccc ctgggagagg gtgggaggag ggagcacaga gactggctgg ggtctcctgg  1900
pCysIlePro TrpGlnArgL euLeuLeuTh rA
gtaggacagg gctgtgagac ggacagaggg ctcctgttgg agcctgaata gggaagagga catcagagag ggacaggagt cacaccagaa aaatcaaatt  2000
gaactggaat tggaaagggg caggaaaacc tcaagagttc tattttccta gttaattgtc actggccact acgtttttaa aaatcataat aactgcatca  2100
gatgacactt taaataaaaa cataaccagg gcatgaaaca ctgtcctcat ccgcctaccg cggacattgg aaaataagcc ccaggctgtg gagggccctg  2200
ggaaccctca tgaactcatc cacaggaatc tgcagcctgt cccaggcact gggtgcaacc aagatcacac aaatccctgc cctcatgaag ctcatgctct  2300
catggggagg aagacagaca tacaaagaga tctagaatgt gaggtcaggt gttgacaaga gcctggaggg aatagagcag ggaaaggtca gaaaaggaag  2400
acccaaggtc tctagaggag gtgtcaggga agggatctcc caagaatgcc ctgatgtgag caggacctga aggcaatggg gagggagccg tgaagacccc  2500
tggaaaagca gattccacac agggaaatgc caaggtcgga ggtgctaagg aaataggaga cacactgctg accttgacct agtaggacac acacacacac  2600
acacacacac actcactcac tccagggctg gggatgaag  agacctgctc aggacccagg accccatttt tccaccctaa tgcataggtc ccaatattga  2700
                           ┌─── Leader                            ┌─── N-domain
ccgatgctct ctgctctctc ctagCCTCAC TTCTAACCTT CTGGAACCCG CCCACCACTG CCAAGCTCAC TATTGAATCC ACGCCGTTCA ATGTCGCAGA  2800
                          laSerL euLeuThrPh eTrpAsnPro ProThrThrA laLysLeuTh rIleGluSer ThrProPheA snValAlaGl
GGGGAAGGAG GTGCTTCTAC TTGTCCACAA TCTGCCCCAG CATCTTTTTG GCTACAGCTG GTACAAAGGT GAAAGAGTGG ATGGCAACCG TCAAATTATA  2900
uGlyLysGlu ValLeuLeuL euValHisAs nLeuProGln HisLeuPheG lyTyrSerTr pTyrLysGly GluArgValA spGlyAsnAr gGlnIleIle
GGATATGTAA TAGGAACTCA ACAAGCTACC CCAGGGCCCG CATACAGTGG TCGAGAGATA ATATACCCCA ATGCATCCCT GCTGATCCAG AACATCATCC  3000
GlyTyrValI leGlyThrGl nGlnAlaThr ProGlyProA laTyrSerGl yArgGluIle IleTyrProA snAlaSerLe uLeuIleGln AsnIleIleG
                                                                                         ┌──Intron 2
AGAATGACAC AGGATTCTAC ACCCTACACG TCATAAAGTC AGATCTTGTG AATGAAGAAG CAACTGGCCA GTTCCGGGTA TACCgtgagt gattccccca  3100
lnAsnAspTh rGlyPheTyr ThrLeuHisV alIleLysSe rAspLeuVal AsnGluGluA laThrGlyGl nPheArgVal TyrP
tgacctctgg gtgttggggg tcagttctac ttcccacaca caggattatc aggcctgggc tgtgctgtgg ccccctctgc attacgaacc atgttagggt  3200
                                                      EcoR1                 Pst1
ttgggcattt agtgcaggat acacacagaa gagacaaact tcaacagatc agaattcctt tccggcatcc agaccctgca g.......(ca 4.6kb)...  7900*
tgttccagat catcgtgcat ctgtcttgtg acacatgcac ccaccgtggg xtttttaagg gctcaggtgg gctgagaggt ggaaggtgcc aactctgatt  8000
                                                                  ┌──A1 domain
gaaagatgcc tgtgaggaat caaaggtxca cacagggxaa tcttctctct gttatctgca agCGGAGCTG CCCAAGCCCT CCATCTCCAG CAACAACTCC  8100
                                                                    roGluLeu ProLysProS erIleSerSe rAsnAsnSer
AAACCCGTGG AGGACAAGGA TGCTGTGGCC TTCACCTGTG AACCTGAGAC TCAGGACGCA ACCTACCTGT GGTGGGTAAA CAATCAGAGC CTCCCGGTCA  8200
LysProValG luAspLysAs pAlaValAla PheThrCysG luProGluTh rGlnAspAla ThrTyrLeuT rpTrpValAs nAsnGlnSer LeuProValS
GTCCCAGGCT GCAGCTGTCC AATGGCAACA GGACCCTCAC TCTATTCAAT GTCACAAGAA ATGACACAGC AAGCTACAAA TGTGAAACCC AGAACCCAGT  8300
erProArgLe uGlnLeuSer AsnGlyAsnA rgThrLeuTh rLeuPheAsn ValThrArgA snAspThrAl aSerTyrLys CysGluThrG lnAsnProVa
                                            ┌──Intron 3
GAGTGCCAGG CGCAGTGATT CAGTCATCCT GAATGTCCTC Tgtgagtata tctgctcctc tctggcccag gctgccagcc aaatccacag ggccagaxxg  8400
lSerAlaArg ArgSerAspS erValIleLe uAsnValLeu T
ggatttctca gtccctctca ggttcaagta cacagaccct caacxctgga catcxagact gtctgtgact ttctgcccag aaaa....(ca 0.4 kb)..  8900*
                                                                                              ┌── B1 domain
ttagagggac actgtggtcc ttccatagac caggaacttc cacttccctc tgacaatatc acctgtggct ttattttgtt tgctccagAT GGCCCGGATG  9000
                                                                                              yr GlyProAspA
CCCCCACCAT TTCCCCTCTA AACACATCTT ACAGATCAGG GGAAAATCTG AACCTCTCCT GCCACGCAGC CTCTAACCCA CCTGCACAGT ACTCTTGGTT  9100
laProThrIl eSerProLeu AsnThrSerT yrArgSerGl yGluAsnLeu AsnLeuSerC ysHisAlaAl aSerAsnPro ProAlaGlnT yrSerTrpPh
TGTCAATGGG ACTTTCCAGC AATCCACCCA AGAGCTCTTT ATCCCCAACA TCACTGTGAA TAATAGTGGA TCCTATACGT GCCAAGCCCA TAACTCAGAC  9200
eValAsnGly ThrPheGlnG lnSerThrGl nGluLeuPhe IleProAsnI leThrValAs nAsnSerGly SerTyrThrC ysGlnAlaHi sAsnSerAsp
                                            ┌── Intron 4
ACTGGCCTCA ATAGGACCAC AGTCACGACG ATCACAGTCT ATGgtaagtg gatccacgaa gcactgacat catgttttga ggtggagtct gtctggtttt  9300
ThrGlyLeuA snArgThrTh rValThrThr IleThrValT yrA
caaacaagag ccaggaagac atttctatc  ccagcctgtg tccagtgggc acaagcaaat cccagattct cccactgaac ctccccaata tgtctctaca  9400
gactcttttc ttcttgttct gatttctcat ggcgggcccc aggccagctt ggaatgtggg gcggaggctc cctcagcccc acagccctgt gtagtggagg  9500
aag....... .......... .......... .....(ca 1.6 kb)..... .......... .......... .......... .......... .......... 11,100*
aggtctctga gccctcagat cattgtgcat ctgtcttgtg acgcacacac acctgccatg agcttttaax gactcagttg ggctgagagg tgggagatgc 11,200
```

```
                                                                                   ┌── A2 domain
caactctgat tgatagatgc ccgtggagga atcacaggtg ccacacaggg caatcttctc tctgttatct gcacagCAGA GCCACCCAAA CCCTTCATCA 11,300
                                                                                   laGl uProProLys ProPheIleT
CCAGCAACAA CTCCAACCCC GTGGAGGATG AGGATGCTGT AGCCTTAACC TGTGAACCTG AGATTCAGAA CACAACCTAC CTGTGGTGGG TAATAATCAG 11,400
hrSerAsnAs nSerAsnPro ValGluAspG luAspAlaVa lAlaLeuThr CysGluProG luIleGlnAs nThrThrTyr LeuTrpTrpV alAsnAsnGl
GAGCCTCCCG GTCAGTCCCA GGCTGCAGCT GTCCAATGAC AACAGGACCC TCACTCTACT CAGTGTCACA AGGAATGATG TAGGACCCTA TGAGTGTGGA 11,500
nSerLeuPro ValSerProA rgLeuGlnLe uSerAsnAsp AsnArgThrL euThrLeuLe uSerValThr ArgAsnAspV alGlyProTy rGluCysGly
                                                            ┌── Intron 5
ATCCAGAACG AATTAAGTGT TGACCACAGC GACCCAGTCA TCCTGAATGT CCTCTgtgag tatcttctgt tcctctgtgg ctcaggctcg gagcccaaat 11,600
IleGlnAsnG luLeuSerVa lAspHisSer AspProValI leLeuAsnVa lLeuT
ccacatagcc aaagt..... .......... .......(ca 0.5 kb).... .......... .......... .......... .......... .......... 12,100*
                                                                                               ┌── B2 domain
aggggxacac tgttgccctt tcacagacca ggagcttccc ctttgctctg atgacattca cctgtggccc tattctcttt gctccagATG GCCCAGACGA 12,200
                                                                                               yrG lyProAspAs
CCCCACCATT TCCCCCTCAT ACACCTATTA CCGTCCAGGG GTGAACCTCA GCCTCTCCTG CCATGCAGCC TCTAACCCAC CTGCACAGTA TTCTTGGCTG 12,300
pProThrIle SerProSerT yrThrTyrTy rArgProGly ValAsnLeuS erLeuSerCy sHisAlaAla SerAsnProP roAlaGlnTy rSerTrpLeu

ATTGATGGGA ACATCCAGCA ACACACACAA GAGCTCTTTA TCTCCAACAT CACTGAGAAG AACAGCGGAC TCTATACCTG CCAGGCCAAT AACTCAGCCA 12,400
IleAspGlyA snIleGlnGl nHisThrGln GluLeuPheI leSerAsnIl eThrGluLys AsnSerGlyL euTyrThrCy sGlnAlaAsn AsnSerAlaS
                                              ┌── Intron 6
GTGGCCACAG CAGGACTACA GTCAAGACAA TCACAGTCTC TGgtaagtgg atccctggac cgttagcaat atgttctgga gcggaatctg tctggttttc 12,500
erGlyHisSe rArgThrThr ValLysThrI leThrValSe rA
agaaaagagc caggaagaaa ttttctttcc tagtatgcat ccaatgggca caagcaatcc caaattcaat cctgagcact cccaatttgt ctctacaaac 12,600
actctt.... .......... .......... .......(ca 1.6 kb)... .......... .......... .......... .......... .......... 14,200*
aagcagccgg ccttacagtc tctgagccct cagatcatcg tacatctgtc ttgtgataca cacacctgcc atgggctttt aaggactcgg gtgggctgaa 14,300
                                                                                                ┌── A3 domain
gggtgggagt tgccaactct gattgaaaga tgcctgtgag gaatcaaaag tgccacacag ggcaatcttc tctctgttat ctgcacagCG GAGCTGCCCA 14,400
                                                                                                la GluLeuProL
AGCCCTCCAT CTCCAGCAAC AACTCCAAAC CCGTGGAGGA CAAGGATGCT GTGGCCTTCA CCTGTGAACC TGAGGCTCAG AACACAACCT ACCTGTGGTG 14,500
ysProSerIl eSerSerAsn AsnSerLysP roValGluAs pLysAspAla ValAlaPheT hrCysGluPr oGluAlaGln AsnThrThrT yrLeuTrpTr
GGTAAATGGT CAGAGCCTCC CAGTCAGTCC CAGGCTGCAG CTGTCCAATG GCAACAGGAC CCTCACTCTA TTCAATGTCA CAAGAAATGA CGCAAGAGCC 14,600
pValAsnGly GlnSerLeuP roValSerPr oArgLeuGln LeuSerAsnG lyAsnArgTh rLeuThrLeu PheAsnValT hrArgAsnAs pAlaArgAla
                                                                         ┌── Intron 7
TATGTATGTG GAATCCAGAA CTCAGTGAGT GCAAACCGCA GTGACCCAGT CACCCTGGAT GTCCTCTgtg agtatctctg ttcctctgtg gccctggttt 14,700
TyrValCysG lyIleGlnAs nSerValSer AlaAsnArgS erAspProVa lThrLeuAsp ValLeuT
.......... .......... .......... .......(ca 1.8 kb)... .......... .......... .......... .......... .......... 16,500*
atttggactt tttaacacag xattgggaca gxattcagag gxacactgtg gcxxttctac aatcaggagc ttcccctttc ctctgatgac atcacctgtg 16,600
                       ┌── B3 domain
gctttgttct ctttgttcca gATGGGCCGG ACACCCCCAT CATTTCCCCC CCAGACTCGT CTTACCTTTC GGGAGCGAAC CTCAACCTCT CCTGCCACTC 16,700
                       yrGlyProA spThrProIl eIleSerPro ProAspSerS erTyrLeuSe rGlyAlaAsn LeuAsnLeuS erCysHisSe
GGCCTCTAAC CCATCCCCGC AGTATTCTTG GGGTATCAAT GGGATACCGC AGCAACACAC ACAAGTTCTC TTAATCGCAA AAATCCAGCC AAATAATAAC 16,800
rAlaSerAsn ProSerProG lnTyrSerTr pArgIleAsn GlyIleProG lnGlnHisTh rGlnValLeu PheIleAlaL ysIleThrPr oAsnAsnAsn
                                                                                   ┌── Intron 8
GGGACCTATG CCTGTTTTGT CTCTAACTTG GCTACTGGCC GCAATAATTC CATAGTCAAG AGCATCACAG TCTCTGgtaa gtggctccct ggagcatcag 16,900
GlyThrTyrA laCysPheVa lSerAsnLeu AlaThrGlyA rgAsnAsnSe rIleValLys SerIleThrV alSerA
                                                                                          NcoI
catcatattc tggggtggag tctatctggt tctcaccaaa gagccaagaa gacattttct ttcccagtct gtgttccatg ggcacaagga aatcccaaat 17,000
tctatcctga gcxccctcac .......... .......(ca 4.8 kb)... .......... .......... .......... .......... .......... 21,800*
                    ┌── M-domain
tccatgacgg acgattcagC ATCTGGAACT TCTCCTGGTC TCTCAGCTGG GGCCACTGCT GGCATCATGA TTGGAGTGCT GGTTGGGGTT GCTCTGATAT 21,900
                    l aSerGlyThr SerProGlyL euSerAlaGl yAlaThrVal GlyIleMetI leGlyValLe uValGlyVal AlaLeuIle*
                                         ┌── Intron 9                                 PstI
AGCAGCCCTG GTGTAGTTTC TTCATTTCAG GAAGACTGgt aggtataatg gcctttcctc ttgttctgtt tcctgcag.. (ca 2.0 kb).......... 24,000*
**                                                      ┌── 3'UT
aattatcatc agatttttaa ctgtactcat tttaaatctt gtcattcaca gACAGTTGTT TTGCTTCTTC CTTAAAGCAT TTGCAACAGC TACAGTCTAA 24,100
                                    Alu repeat
AATTGCTTCT TTACCAAGGA TATTTACAGA AAAGACTCTG ACCAGAGATC GAGACCATCC TAGCCAACAT CGTGAAACCC CATCTCTACT AAAAATACAA 24,200
AAATGAGCTG GGCTTGGTGG CGCGCACCTG TAGTCCCAGT TACTCGGGAG GCTGAGGCAG GAGAATCGCT TGAACCCGGG AGGTGGAGAT TGCAGTGAGC 24,300
CCAGATCGCA CCACTGCACT CAGTCTGGCA ACAGAGCAAG ACTCCATCTC AAAAAGAAAA GAAAAGAAGA CTCTGACCTG TACTCTTGAA TACAAGTTTC 24,400
TGATACCACT GCACTGTCTG AGAATTTCCA AAACTTTAAT GAACTAACTG ACAGCTTCAT GAAACTGCCA CCAAGATCAA GCAGAGAAAA TAATTAATTT 24,500
CATGGGACTA AATGAACTAA TGAGGATAAT ATTTTCATAA TTTTTTATTT GAAATTTTGC TGATTCTTTA AATGTCTTGT TTCCCAGATT TCAGGAAACT 24,600
TTTTTTCTTT TAAGCTATCC ACAGCTTACA GCAATTGATA AAATATACTT TTGTGAACAA AAATTGAGAC ATTTACATTT TCTCCCTAGT GGTCGCTCCA 24,700
                                                                    Poly A
GACTTGGGAA ACTATTCATG AATATTTATA TTGTATGGTA ATATAGTTAT TGCACAAGTT CAATAAAAAT CTGCTCTTTG TATGACAGAA TACATTTGAA 24,800
                                                                           Poly A
AACATTGGTT ATATTACCAA GACTTTGACT AGAATGTCGT ATTTGAGGAT ATAAACCCAT AGGTAATAAA CCCACAGGTA CTACAAACAA AGTCTGAAGT 24,900
                                                                                                  Alu repeat
CAGCCTTGGT TTGGCTTCCT AGTGTCAATT AAACTTCTAA AAGTTTAATC TGAGATTCCT TATAAAAACT TCCAGCAAAG CAACTTTAAA AAAGTCTGTG 25,000
TGGGCCGGGC GCGGTGGCTC ACGCCTGTAA TCCCAGCACT TTCTACCGCC GAGG.....(ca 180 bp Alu repeat)  .......... .......... 25,200*
AGTGAGCCAA GATACTGCGC TGCACTCCAG CCTGXGAGAC AAAGTGAGAC TCCGTCAAAA AAAAAAAAAA GTCTATGTGG TCAGTCACTA CTCTTGCCTG 25,300
                                     Poly A
CAGTTATGAA AAGAATGAGG CCAAGTTGAT GAAAATAAAC TTATTTTGAA AACAAATTTG TCTTATTGGG ATTATCTTTG GTAGAAATAG AGATGCCTAT 25,400
                                                                                                   ┌── "NCA-like" exon
GAAGAGAAAT TATGTTGAAA AAAAACTATA GTACACCTGT TATGAGACTG TCACTTTGTA CATTGTTGAG TTTTTATTAT CCACCTGTAG ACTAGAGTGG 25,500
    EcoR1
ACCATGAATT C
```

- 1 Leader -22 1 N-terminal Domain
MESPSAPPHRWCIPWQRLLLTASLLTFWNPPTTA KLTIESTP......

108 A1 201 B1
EATGQFRVYPEL**P**K**P**S...............VILNVLYG**P**DA**P**.....

285 A2 379 B2
TTVTTITVYAE**PP**K**P**F...............VILNVLYG**P**DD**P**.....

464 A3 557 B3
TTVKTITVSAEL**P**K**P**S...............VTLDVLYG**P**DT**P**I.....

642 M-domain 668
SIVKSITVSASGTSPGLSAGATVGIMIGVLVGVALI

FIG. 3. Correlation of exons with protein domains in CEA. The amino acid sequence, including the leader and membrane peptides, is shown in single-letter code over the regions of exon/intron boundaries. The amino acids whose codons are split by exon/intron junctions are shown in outline. The N-terminal leader peptide is split at codon −22, a common feature in immunoglobulin-related genes. The N terminus of the mature protein is lysine (K). Each of the immunoglobulin-like domains (N-terminal, A1, B1, A2, B2, A3, B3) ends with a predicted β-strand (underlined) and begins (except for the N-terminal domain) with a predicted interconnecting loop, rich in proline residues (bold). The membrane peptide (M-domain), which serves as a signal for attachment of the phosphatidylinositol glycan moiety, is at the beginning of a separate exon.

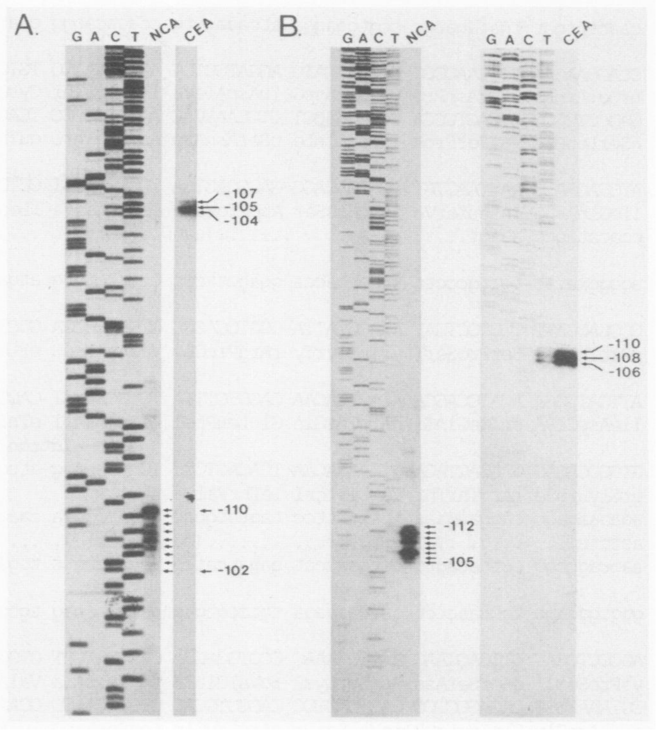

FIG. 4. Determination of the transcriptional starts of the NCA and CEA genes through primer extension (A) and S1 nuclease (B) analyses. As length markers, DNA sequences of the corresponding genomic regions, with the same NCA (A) or NCA and CEA (B) oligonucleotide primers for each gene as for the transcriptional start analyses (NCA5′ and CEAL, respectively), were used to allow the start positions to be determined exactly. In both cases, the numbers indicate the transcriptional start sites with respect to the initiation codons of each gene.

activity in both cell lines. These values ranged from 4 to 20% of the pSV2CAT activity. The CEA constructs, however, exhibited a much higher CAT expression in the SW403 cells, in which the mean value was approximately 110% for the pCEA424/2CAT construct, compared with the HeLa cells, in which the values ranged from 5 to 12% of the pSV2CAT activity for all CEA constructs. The expression levels of the longer CEA constructs (pCEA832/2CAT and pCEA3300/2CAT) were also elevated in SW403 cells compared with HeLa cells (50 and 40% of the pSV2CAT activity, respectively). The constructs for both genes in which the promoters were in the inverse orientation (pNCA48/1246CAT and pCEA2/832CAT) showed no promoter activity (<1%) in either cell line. The pBLCAT3 vector, which lacks a promoter entirely, was also not active (<1%).

DISCUSSION

We have isolated a cosmid clone (cosCEA1) from a human genomic library; this clone contains the complete coding sequence for CEA, as determined by restriction endonuclease mapping and DNA sequencing (Fig. 1 and 2). Another group recently reported the independent isolation and characterization of an identical clone from the same genomic library (T. C. Willcocks and I. W. Craig, Abstr. XVIIth Meet. Int. Soc. Oncodevel. Biol. Med. 1989, abstr. no. 1P-1, p. 51). The exon structure is the same as that reported for other members of the CEA gene family for the immunoglobulin variable- and constant-like domains (33–35, 50, 51) and, likewise, for other members of the immunoglobulin supergene family (58). All of the immunoglobulin-like domain exon sequences reported here are present in the mature CEA mRNA, whose sequence has been determined elsewhere (2, 4, 36, 59). Hybridization studies with restriction endonuclease fragments from the N-terminal and repeat domains of a CEA cDNA do not indicate the existence of additional immunoglobulin variable- or constant-like domain exons in the CEA gene.

Analysis of the 3′ end of the gene has revealed that the hydrophobic membrane domain is encoded by a separate exon (M/3′), which also contains the first 39 nucleotides of the 3′ untranslated region (Fig. 1 and 2). This region is homologous to a corresponding section of the 3′ untranslated region found in NCA cDNA sequences (2, 31, 47). After this, the sequences of the 3′ untranslated regions of the CEA and NCA mRNAs are no longer homologous. The point of their divergence correlates exactly with the end of this M/3′ exon in the CEA gene (Fig. 2). The following exon in the CEA gene contains the rest of the CEA 3′ untranslated region (Fig. 1 and 2, 3′ exon). Preliminary characterization of the region directly downstream from this exon has revealed a sequence which shows homology to the NCA 3′ untranslated region ("NCA" 3′ exon), including a conserved *Eco*RI restriction endonuclease site (Fig. 2; data not shown). A consensus intron acceptor site (30) exists at the 5′ end of this NCA-like exon (Fig. 2), which correlates with a corresponding splice site in a related PSG gene (33). This PSG gene also contains both CEA and NCA-like 3′ untranslated regions, thus confirming the model, originally proposed by Oikawa et al. (33) and extended by Zimmermann et al. (61), that all genes belonging to the CEA family contain such a complex 3′ unit. Therefore, variability in the 3′ untranslated regions in the CEA and NCA mRNAs is probably due to differential splicing of their corresponding genes. These results also indicate that a CEA transcript which contains the NCA-like 3′ untranslated region could exist and that the probe from the

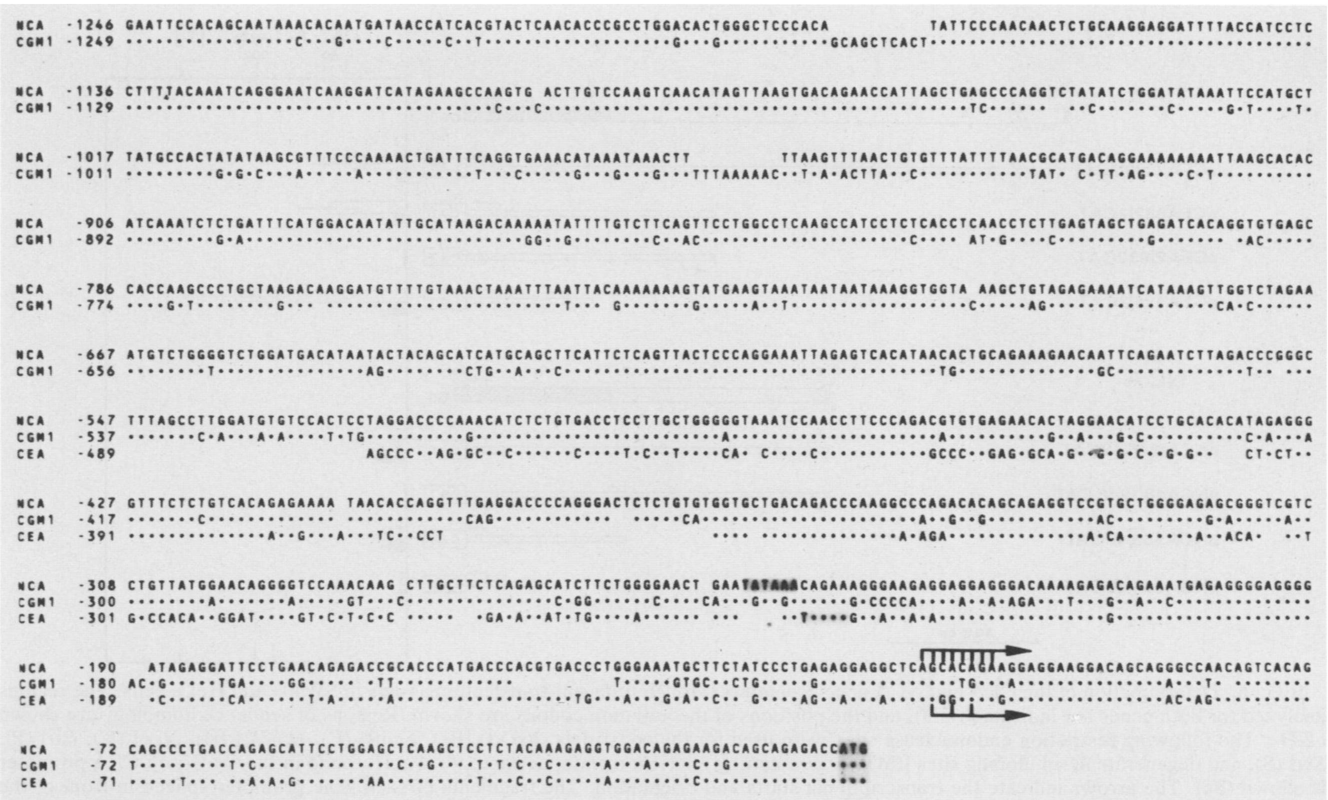

FIG. 5. Comparison of the nucleotide sequences from the putative promoter regions of the NCA, CEA, and CGM1 genes. The numbers indicate the distance in nucleotides from the initiation codon for each gene. Gaps have been introduced to allow optimal alignment. Identical nucleotides are indicated by dots. The cluster of transcriptional start sites determined for CEA and NCA in the S1 nuclease assays (see Fig. 4 and text) are designated (arrows). Consensus TATA sequences for the NCA and CEA genes, as well as the initiation codons, are shaded grey.

3′ untranslated region of an NCA cDNA clone (60) may not be NCA mRNA specific.

Recently, we found that after stable transfection of this CEA cosmid clone into Chinese hamster ovary (CHO) cells, a 180,000-molecular-weight glycosylated CEA molecule was expressed on the cell surface (L. J. F. Hefta, H. Schrewe, J. A. Thompson, S. Oikawa, H. Nakazato, and J. E. Shively, Cancer Res., in press). These results show that a functional promoter is probably present in the 3,300-nucleotide region upstream of the translational start contained within this cosmid clone (Fig. 1). Although a rapid sequence divergence between the human and rodent CEA families has been reported, which indicates a parallel but independent evolution of the genes in different mammalian orders (39), these expression results suggest that the *trans*-acting factors and *cis*-acting elements responsible for activating the CEA gene are conserved enough between humans and rodents to allow expression across species. Therefore, closer analysis of the regulation of human CEA gene expression in a rodent system is feasible. Furthermore, the encoded CEA molecule was apparently posttranslationally modified through addition of a phosphatidylinositol glycan tail, as for the native CEA molecule (Hefta et al., in press). These results indicate that all the information needed for the correct expression of the CEA gene is contained within cosCEA1.

To define the actual portion of the 5′ untranslated region which is required for the promoter activity of the CEA gene and, for comparison, that of a closely related NCA gene, we carried out functional tests by placing restriction endonuclease fragments of various lengths from the putative promoter regions of both genes upstream of the CAT reporter gene and assaying for CAT activity in a transient transfection assay in two different human cell lines (Fig. 6). For this purpose, we chose the CEA-producing adenocarcinoma cell line SW403 and, as a negative control, the HeLa cell line. No significant differences in the promoter activities could be determined between the cell lines with the NCA constructs used. These seem to contain only a basal level of promoter activity, especially for the shortest construct, which showed minimal activity (Fig. 6). On the other hand, the CEA promoter constructs showed an enhanced expression of the CAT gene in SW403 cells, which was nine times greater than in HeLa cells, when the shortest construct was used (Fig. 6, pCEA424/2CAT). It appears that *cis* regulatory sequences, which are responsible for this enhancement, along with a functional transcription initiator, are both present within the first 424 nucleotides upstream of the translational start. When the sequences in this region of the CEA and NCA genes were compared, homology was found (Fig. 5 and 6); however, regions showing stronger sequence divergence also exist in the region upstream of −240 nucleotides from the translational start of the CEA gene, which may be of interest regarding the differential regulation of expression of the two genes. Further experiments must be designed to analyze these regions more closely. It is also interesting that the longer CEA constructs are approximately 50% less active in their promoter activities in SW403 cells and 25% less active in HeLa cells than is the pCEA424/2CAT construct (Fig. 6). A possible explanation for this phenomenon is that a silencer region could exist between nucleotides

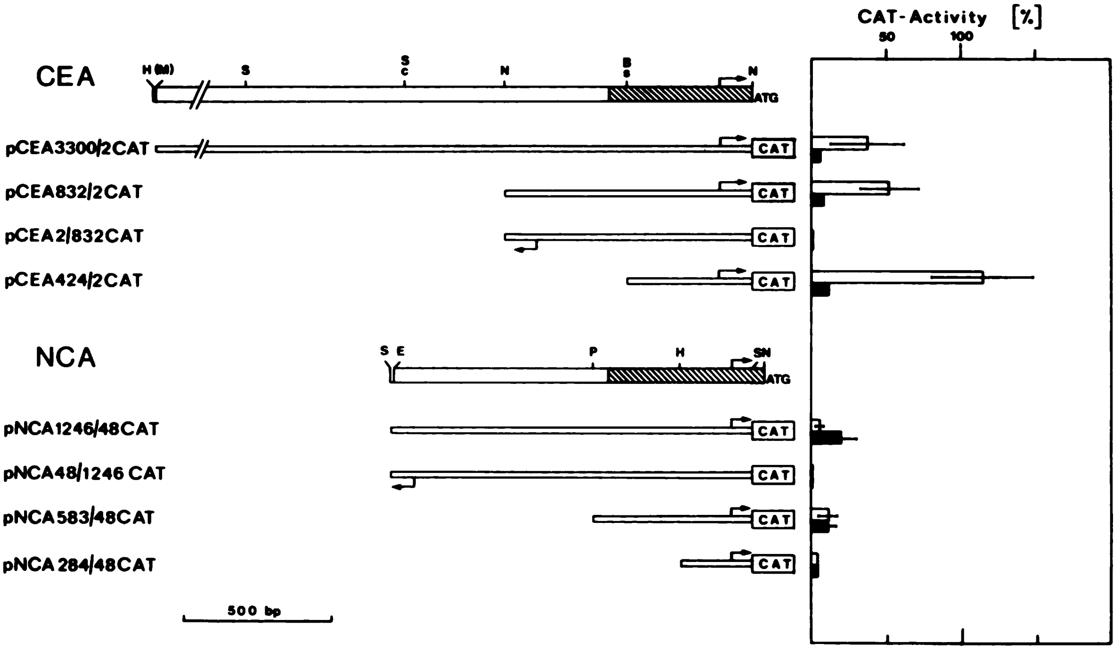

FIG. 6. Determination of the CEA and NCA gene promoters through transient-transfection assays in SW403 and HeLa cells. The regions analyzed for both genes are indicated (□), and the positions of the initiation codons are shown. Regions of sequence homology are shown (▨). The following restriction endonuclease sites were used for the constructs: *Bst*XI, (Bs), *Eco*RI (E), *Hin*dIII (H), *Nco*I (N), *Pst*I (P), *Sst*I (S), and degenerate *Msp*I cloning sites [(M)]. A rare-cutting restriction endonuclease site (*Sac*II) found in the C+G-rich CEA promoter is shown (Sc). The arrows indicate the transcriptional starts and orientation. The fragments of each gene promoter spliced in front of the promoterless CAT reporter gene are shown (▭). The CAT activity was determined for each construct after transient transfection into SW403 cells (□) and HeLa cells (■). The activities are expressed as a percentage of the CAT activity obtained with the plasmid pSV2CAT (see text), which was equivalent to a mean of 30,000 dpm in SW403 or 50,000 dpm in HeLa cells (ca. 30 and 50% conversion of the acetyl coenzyme A added, respectively; this was the limiting factor). Up to five separate assays were carried out for each construct to confirm reproducibility. Standard deviations are indicated with bars.

−424 and −832 upstream from the translational start, which reduces the activities in both cell lines through interaction with common *trans*-acting regulatory factors. Such silencer sequences have indeed been described for other genes (8, 9). However, further studies must be carried out to analyze this in more detail.

As found here for CEA and NCA, a number of other eucaryotic genes have also been reported which do not contain obvious TATA boxes. The promoters of such genes can be divided into two classes (42). The members of the first class are G+C rich and are found primarily in housekeeping genes (41). These promoters usually contain several transcription initiation sites spread over a fairly large region, as well as potential binding sites for Spl (11). The members of the second class are not G+C rich, are not constitutively active, but are regulated during differentiation or development and initiate transcription at only one or a few tightly clustered start sites (42). Included in this class are a number of genes that are regulated during mammalian immunodifferentiation, e.g., the T-cell receptor β-chain genes (1) and the V_{preB} gene (22), as well as some *Drosophila* homeotic genes (6, 38, 43). The CEA and NCA genes show a closer resemblance to this latter group, because their promoters are not obviously G+C rich, they contain no identifiable Spl-binding sites, they reveal only a few tightly clustered start sites, and, most importantly, they are not constitutively expressed.

Future experiments will concentrate on more closely defining the various regions of interest in the promoters and, in parallel, will be directed toward identifying regions which possibly convey an oncofetally regulated expression pattern of the CEA gene.

ACKNOWLEDGMENTS

We thank Kurt Lucas for his help with the computer analyses and Sabine Barnert for her excellent technical assistance. Special thanks to W. Lindenmaier for the cosmid library.

This work was supported by the Deutsche Forschungsgemeinschaft and Public Health Service grant CA-37808 from the National Cancer Institute.

LITERATURE CITED

1. **Anderson, S. J., H. S. Chou, and D. Y. Loh.** 1988. A conserved sequence in the T-cell receptor β-chain promoter region. Proc. Natl. Acad. Sci. USA **85:**3551–3554.
2. **Barnett, T., S. J. Goebel, M. A. Nothdurft, and J. J. Elting.** 1988. Carcinoembryonic antigen family: characterization of cDNAs coding for NCA and CEA and suggestion of nonrandom sequence variation in their conserved loop-domains. Genomics **3:**59–66.
3. **Barnett, T. R., A. Kretschmer, D. A. Austen, S. J. Goebel, J. T. Hart, J. J. Elting, and M. E. Kamarck.** 1989. Carcinoembryonic antigens: alternative splicing accounts for the multiple mRNAs that code for novel members of the carcinoembryonic antigen family. J. Cell Biol. **108:**267–276.
4. **Beauchemin, N., S. Benchimol, D. Cournoyer, A. Fuks, and C. F. Stanners.** 1987. Isolation and characterization of full-length functional cDNA clones for human carcinoembryonic antigen. Mol. Cell. Biol. **7:**3221–3230.
5. **Benchimol, S., A. Fuks, S. Jothy, N. Beauchemin, K. Shirota, and C. P. Stanners.** 1989 Carcinoembryonic antigen, a human tumor marker, functions as an intercellular adhesion molecule.

Cell **57**:327–334.

6. **Biggin, M. D., and R. Tjian.** 1988. Transcription factors that activate the Ultrabithorax promoter in developmentally staged extracts. Cell **53**:699–711.
7. **Boucher, D., D. Cournoyer, C. P. Stanners, and A. Fuks.** 1989. Studies on the control of gene expression of the carcinoembryonic antigen family in human tissue. Cancer Res. **49**:847–852.
8. **Brand, A. M., L. Breeden, J. Abraham, R. Sternglanz, and K. Nasmyth.** 1985. Characterization of a "silencer" in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer. Cell **41**:41–48.
9. **Cao, S. X., P. D. Gutman, H. P. G. Dave, A. Schechter.** 1989. Identification of a transcriptional silencer in the 5′-flanking region of the human ε-globin gene. Proc. Natl. Acad. Sci. USA **86**:5306–5309.
10. **Cordon, J., B. Wasylyk, A. Buchwalder, P. Sassone-Corsi, C. Kedinger, and P Chambon.** 1980. Promoter sequences of eukaryotic protein-coding genes. Science **209**:1406–1414.
11. **Dynan, W. S., and R. Tjian.** 1983. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. Cell **35**:79–87.
12. **Felgner, P. L., T. R. Gadek, M. Holm, R. Roman, H. W. Chan, M. Wenz, J. P. Northrop, G. M. Ringold, and M. Danielsen.** 1987. Lipofection: a highly efficient, lipid-mediated DNA-transfection procedure. Proc. Natl. Acad. Sci. USA **84**:7413–7417.
13. **Gold, P., and S. O. Freedman.** 1965. Demonstration of tumor-specific antigens in human colonic carcinomata by immunological tolerance and absorption techniques. J. Exp. Med. **121**: 439–462.
14. **Gorman, C. M., L. F. Moffat, and B. H. Howard.** 1982. Recombinant genomes which express chloramphenicol acetyltransferase in mammalian cells. Mol. Cell. Biol. **2**:1044–1051.
15. **Grunert, F., F. Kolbinger, K. Schwarz, H. Schwaibold, and S. von Kleist.** 1988. Protein analysis of NCA-50 shows identity to NCA cDNA deduced sequences and indicates posttranslational modifications. Biochem. Biophys. Res. Commun. **153**:1105–1115.
16. **Hefta, S. A., L. J. F. Hefta, T. D. Lee, R. J. Paxton, and J. E. Shively.** 1988. Carcinoembryonic antigen is anchored to membranes by covalent attachment to a glycosylphosphatidylinositol moiety: identification of the ethanolamine linkage site. Proc. Natl. Acad. Sci. USA **85**:4648–4652.
17. **Hinoda, Y., M. Neumaier, S. A. Hefta, Z. Drzenick, C. Wagener, L. Shively, L. J. Hefta, J. E. Shively, and R. J. Paxton.** 1988. Molecular cloning of a cDNA coding biliary glycoprotein I: primary structure of a glycoprotein immunologically crossreactive with carcinoembryonic antigen. Proc. Natl. Acad. Sci. USA **85**:6959–6963. (Erratum, **86**:1668, 1989.)
18. **Hümbelin, M., B. Safer, J. A. Chiorini, J. W. B. Hershey, and R. B. Cohen.** 1989. Isolation and characterization of the promoter and flanking regions of the gene encoding the human protein-synthesis-initiation factor 2α. Gene **81**:315–324.
19. **Jean, F., P. Malapert, G. Rougon, and J. Barbet.** 1988. Cell membrane, but not circulating, carcinoembryonic antigen is linked to a phosphatidylinositol-containing hydrophobic domain. Biochem. Biophys. Res. Commun. **155**:794–800.
20. **Kodelja, V., K. Lucas, S. Barnert, S. von Kleist, J. A. Thompson, and W. Zimmermann.** 1989. Identification of a carcinoembryonic antigen gene family in the rat: analysis of the N-terminal domains reveals immunoglobulin-like, hypervariable regions. J. Biol. Chem. **264**:6906–6912.
21. **Kolbinger, F., K. Schwarz, F. Brombacher, S. von Kleist, and F. Grunert.** 1989. Expression of an NCA cDNA in NIH/3T3 cells yields a 110K glycoprotein, which is anchored into the membrane via glycosyl-phosphatidylinositol. Biochem. Biophys. Res. Commun. **161**:1126–1134.
22. **Kudo, A., and F. Melchers.** 1987. A second gene, V_{preB} in the λ_5 locus of the mouse, which appears to be selectively expressed in pre-B lymphocytes. EMBO J. **6**:2267–2272.
23. **Kuroki, M., F. Arakawa, H. Yamamoto, Y. Ikehara, and Y. Matsuoka.** 1988. Active production and membrane anchoring of carcinoembryonic antigen observed in normal colon mucosa. Cancer Lett. **43**:151–157.
24. **Lathe, R.** 1985. Synthetic oligonucleotide probes deduced from amino acid sequence data: theoretical and practical considerations. J. Mol. Biol. **183**:1–12.
25. **Leibovitz, A., J. C. Stinson, W. B. McCombs, E. Cameron, K. C. Mazur, and N. D. Mabry.** 1976. Classification of human colorectal adenocarcinoma cell lines. Cancer Res. **36**:4562–4569.
26. **Lin, T.-M., and S. P. Halpert.** 1976. Placental localization of human pregnancy-associated plasma proteins. Science **193**: 1249–1252.
27. **Lindenmaier, W., H. Hauser, I. Greiser de Wilke, and G. Schütz.** 1982. Gene shuttling: moving of cloned DNA into and out of eukaryotic cells. Nucleic Acids Res. **10**:1243–1256.
28. **Luckow, B., and G. Schütz.** 1987. CAT constructions with multiple unique restriction sites for the functional analysis of eukaryotic promoters and regulatory elements. Nucleic Acids Res. **15**:5490.
29. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
30. **Mount, S. M.** 1982. A catalogue of splice junction sequences. Nucleic Acids Res. **10**:459–472.
31. **Neumaier, M., W. Zimmermann, L. Shively, Y. Hinoda, A. D. Riggs, and J. E. Shively.** 1988. Characterization of a cDNA clone for the nonspecific crossreacting antigen (NCA) and a comparison of NCA and carcinoembryonic antigen (CEA). J. Biol. Chem. **263**:3202–3207.
32. **Oikawa, S., S. Imajo, T. Noguchi, G. Kosaki, and H. Nakazato.** 1987. The carcinoembryonic antigen (CEA) contains multiple immunoglobulin-like domains. Biochem. Biophys. Res. Commun. **144**:634–642.
33. **Oikawa, S., C. Inuzuka, G. Kosaki, and H. Nakazato.** 1988. Exon-intron organization of a gene for pregnancy-specific β1-glycoprotein, a subfamily member of CEA-family: implications for its characteristic repetitive domains and N-terminal sequences. Biochem. Biophys. Res. Commun. **156**:68–77.
34. **Oikawa, S., C. Inuzuka, M. Kuroki, Y. Matsuoka, G. Kosaki, and H. Nakazato.** 1989. A pregnancy-specific β1-glycoprotein, a CEA family member, expressed in a human promyelocytic leukemia cell line, HL-60: structures of protein, mRNA and gene. Biochem. Biophys. Res. Commun. **163**:1021–1031.
35. **Oikawa, S., G. Kosaki, and H. Nakazato.** 1987. Molecular cloning of a gene for a member of carcinoembryonic antigen (CEA) gene family: signal peptide and N-terminal domain sequences of nonspecific crossreacting antigen (NCA). Biochem. Biophys. Res. Commun. **146**:464–469.
36. **Oikawa, S., H. Nakazato, and G. Kosaki.** 1987. Primary structure of human carcinoembryonic antigen (CEA) deduced from cDNA sequence. Biochem. Biophys. Res. Commun. **142**:511–518.
37. **Paxton, R., G. Mooser, H. Pande, T. D. Lee, and J. E. Shively.** 1987. Sequence analysis of carcinoembryonic antigen: identification of glycosylation sites and homology with the immunoglobulin supergene family. Proc. Natl. Acad. Sci. USA **84**: 920–924.
38. **Perkins, K. K., G. M. Dailey, and R. Tjian.** 1988. In vitro analysis of the Antennapedia P2 promoter: identification of a new Drosophila transcription factor. Genes Dev. **2**:1615–1626.
39. **Rudert, F., W. Zimmermann, and J. Thompson.** 1989. Intra- and interspecies analyses of the carcinoembryonic antigen (CEA) gene family reveal an independent evolution in primates and rodents. J. Mol. Evol. **29**:126–134.
40. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
41. **Seghal, A., N. Patil, and M. Chao.** 1988. A constitutive promoter directs expression of the nerve growth factor receptor gene. Mol. Cell. Biol. **8**:3160–3167.
42. **Smale, S. T., and D. Baltimore.** 1989. The "initiator" as a transcription control element. Cell **57**:103–113.
43. **Soeller, W. C., S. J. Poole, and T. Kornberg.** 1988. In vitro transcription of the Drosophila engrailed gene. Genes Dev. **2**:68–81.
44. **Struhl, K.** 1989. Molecular mechanisms of transcriptional regu-

lation in yeast. Annu. Rev. Biochem. **58**:1051–1077.

45. **Takami, N., Y. Misumi, M. Kuroki, Y. Matsuoka, and Y. Ikehara.** 1988. Evidence for carboxyl-terminal processing and glycolipid-anchoring of human carcinoembryonic antigen. J. Biol. Chem. **263**:12716–12720.
46. **Tatarinov, Y. S.** 1978. Trophoblast-specific beta1-glycoprotein as a marker for pregnancy and malignancies. Gynecol. Obstet. Invest. **9**:65–97.
47. **Tawaragi, Y., S. Oikawa, Y. Matsuoka, G. Kosaki, and H. Nakazato.** 1988. Primary structure of nonspecific crossreacting antigen (NCA), a member of carcinoembryonic antigen (CEA) gene family, deduced from cDNA sequence. Biochem. Biophys. Res. Commun. **150**:89–96.
48. **Thompson, J., S. Barnert, B. Berling, S. von Kleist, V. Kodelja, K. Lucas, E.-M. Mauch, F. Rudert, H. Schrewe, M. Weiss, and W. Zimmermann.** 1989. Structure, expression and evolution of the human and rat carcinoembryonic antigen (CEA) gene families, p. 65–74. *In* A. Yachi and J. Shively (ed.), The carcinoembryonic antigen gene family. Elsevier Science Publishers BV, Amsterdam.
49. **Thompson, J., and W. Zimmermann.** 1988. The carcinoembryonic antigen gene family: structure, expression and evolution. Tumor Biol. **9**:63–83.
50. **Thompson, J. A., E.-M. Mauch, F.-S. Chen, Y. Hinoda, H. Schrewe, B. Berling, B. Barnert, S. von Kleist, J. E. Shively, and W. Zimmermann.** 1989. Analysis of the size of the carcinoembryonic antigen (CEA) gene family: isolation and sequencing of N-terminal domain exons. Biochem. Biophys. Res. Commun. **158**:996–1004.
51. **Thompson, J. A., H. Pande, R. J. Paxton, L. Shively, A. Padma, R. L. Simmer, C. T. Todd, A. D. Riggs, and J. E. Shively.** 1987. Molecular cloning of a gene belonging to the carcinoembryonic antigen gene family and discussion of a domain model. Proc. Natl. Acad. Sci. USA **84**:2965–2969.
52. **Thomson, D. M. P., J. Krupey, S. O. Freedman, and P. Gold.** 1969. The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. Proc. Natl. Acad. Sci. USA **64**:161–167.
53. **Toribara, N. W., T. L. Sack, J. R. Gum, S. B. Ho, J. E. Shively, J. K. V. Willson, and Y. S. Kim.** 1989. Heterogeneity in the induction and expression of carcinoembryonic antigen-related antigens in human colon cancer cell lines. Cancer Res. **49**: 3321–3327.
54. **Tran, R., S. V. S. Kashmiri, J. Kantor, J. W. Greiner, S. Pestka, J. E. Shively, and J. Schlom.** 1988. Correlation of DNA hypomethylation with expression of carcinoembryonic antigen in human carcinoma cells. Cancer Res. **48**:5674–5679.
55. **Wallace, R. B., and C. G. Miyada.** 1987. Oligonucleotide probes for the screening of recombinant DNA libraries. Methods Enzymol. **152**:432–442.
56. **Watanabe, S., and J. Y. Chou.** 1988. Human pregnancy-specific β1-glycoprotein: a new member of the carcinoembryonic antigen gene family. Biochem. Biophys. Res. Commun. **152**:762–768.
57. **Williams, A. F.** 1987. A year in the life of the immunoglobulin superfamily. Immunol. Today **8**:298–303.
58. **Williams, A. F., and A. N. Barclay.** 1988. The immunoglobulin superfamily—domains for cell surface recognition. Annu. Rev. Immunol. **6**:381–405.
59. **Zimmermann, W., B. Ortlieb, R. Friedrich, and S. von Kleist.** 1987. Isolation and characterization of cDNA clones encoding the human carcinoembryonic antigen reveal a highly conserved repeating structure. Proc. Natl. Acad. Sci. USA **84**:2960–2964.
60. **Zimmermann, W., B. Weber, B. Ortlieb, F. Rudert, W. Schempp, H.-H. Fiebig, J. E. Shively, S. von Kleist, and J. A. Thompson.** 1988. Chromosomal localization of the carcinoembryonic antigen gene family and differential expression in various tumors. Cancer Res. **48**:2550–2555.
61. **Zimmermann, W., M. Weiss, and J. A. Thompson.** 1989. cDNA cloning demonstrates the expression of pregnancy-specific glycoprotein genes, a subgroup of the carcinoembryonic antigen gene family, in fetal liver. Biochem. Biophys. Res. Commun. **163**:1197–1209.