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

HEINRICH SCHREWÉ, JOHN THOMPSON, MARION BONA, LAURA J. F. HFEITA, AIKO MARUYA, MARTIN HASSAUER, JOHN E. SHIVELY, SABINE VON KLEIST, AND WOLFGANG ZIMMERMANN

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

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 cosm id 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 oncosetal 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 embryoid and tumor development, it is necessary to characterize the CEA gene. In this paper, we report the isolation of a cosm id 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, RNAasin, 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. The cosCEAl DNA fragment from pNCA1 (20) and the CEAL oligonucleotide (see above) were hybridized with DNA from the 3' untranslated region of a CEA cDNA clone.

**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 pHC79-2-cos/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 MgSO4, 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 CEAL-specific, 32P-labeled, 408-base-pair (bp) RsaI-PstI restriction endonuclease fragment from the 3' untranslated region of a CEA cDNA clone (pCEAA3 [60]) was performed in the presence of 50% formamide, 6X SSPE (1X SSPE is 0.18 M NaCl, 10 mM NaH2PO4 [pH 7.4], and 1 mM EDTA), 1X 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 2X 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 32P-labeled synthetic oligonucleotide (CEAL, 5'-AGCCTCTGCGAGGGATGCACT-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 2X 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 AatII, which has three recognition sites in the vector. A 32.6-kilobase-pair (kb) AatII 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 (BamHI, SalI, Smal, SstI, and XbaI), 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'-TAGGGCTAT CAGAGGGCTTTTGC-3'). After stripping, the blot was rehybridized with an oligonucleotide (cos2, 5'-GGCGATGC TGTGGAATGTGAC-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 AatII 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 PstI 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 using oligonucleotide probes homologous to regions at the beginning and end of each A and B immunoglobulin constant-like domains. The cosCEAl 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 5X SSC (1X SSC is 0.15 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 RsaI-EcoRI 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 dyeoxy-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 reverse transcription 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'-TCTCTGTCACACTTCTG TAGAGCA-3', corresponding to positions 15 to 39 in Fig. 5). For the annealing reaction, 5 to 10 ng of oligonucleotide labeled with 32P 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 MgCl2, 1.75 mM dithiothreitol, 3.5 mM deoxyribonucleoside triphosphates, 1 ng of dactinomycin per µl, 2 U of RNAasin per µl, and 0.8 U of reverse transcriptase per µl in a volume of 25 µl.
and incubated at 42°C for 1 h. Following phenol extraction and ethanol precipitation, the extension products were de-natured and loaded onto a 6% polyacrylamide–7 M urea DNA sequencing gel.

(ii) 5' nuclease analyses. The transcriptional starts of CEA and NCA were independently determined by 5' nuclease mapping. For CEA, a 1.6-kb Psrl restriction endonuclease fragment from cosCEA1, containing the first exon of the CEA gene (Fig. 1), and for NCA a 606-bp Psrl 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 Smal (for CEA) or Psrl (for NCA), and the single-stranded 456-nt CEA-specific fragments and the 562-nt 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 colonc tumor (see above) and 5' 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 5' nuclease per ml for 30 min at 18°C and analyzed as above.

Construction of promoter-CAT gene hybrids. The 3.4-kb SsiI–EcoRI fragment of the cosmId clone cosCEA1 (Fig. 1), which contains the first two exons of the CEA gene, was digested with NcoI. The 835-bp NcoI fragment was isolated and treated with 5' 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 BamHI. The ends filled in with the Klenow fragment of E. coli DNA polymerase I, and the blunt-ended NcoI fragment was inserted to generate the constructs pCEA832/2CAT and pCEA2/832CAT (see Fig. 6). A 409-bp Psrl–BsrXI fragment was removed from clone pCEA823/2CAT, 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 HindIII–BsrXI fragment of plasmid pCEA832/2CAT was removed, and the rest was ligated to a 2.9-kb HindIII–BsrXI fragment from the 5' end of cosCEA1 (Fig. 1), generating the construct pCEA3300/2CAT (see Fig. 6).

The 2.7-kb EcoRI 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 SstI, which cuts in the poly linker 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 Psrl, 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 HindIII 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 10⁴ 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 [14C]acetate 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 radioactive 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 cosmids 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 polyuracil sequence was noted at the beginning of the gene (1516 to 1588).
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 (cosCEAI) 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. XVIIIth 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 EcoRI 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
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 these 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...
−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 Sp1 (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 Sp1-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 oncofetaly 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

THE CARCINOEMBRYONIC ANTIGEN GENE

rectal adenocarcinoma cell lines. Cancer Res. 36:4562-4569.
44. Struhl, K. 1989. Molecular mechanisms of transcriptional regu-


