Isolation and characterization of cDNA clones encoding the human carcinoembryonic antigen reveal a highly conserved repeating structure

(tumor antigen/oligonucleotide probes/repeated domain/*Alu* sequence/colon tumor)

WOLFGANG ZIMMERMANN, BARBARA ORTLIEB, ROLAND FRIEDRICH, AND SABINE VON KLEIST

Institut für Immunbiologie, Universität Freiburg, Stefan-Meier-Strasse 8, D-7800 Freiburg, Federal Republic of Germany

ABSTRACT   For the isolation of cDNA clones encoding the carcinoembryonic antigen (CEA), we have constructed a cDNA library from human colon tumor mRNA. The library was screened with various oligonucleotides whose sequence had been deduced from partial amino acid sequence data for CEA. Positive candidate clones were hybridized with a probe for repetitive DNA, because CEA mRNA contains an *Alu* repetitive element, and with a fragment of a genomic clone of nonspecific cross-reacting antigen, an antigen closely related to CEA. Here we report the nucleotide sequence of the two overlapping CEA cDNA clones comprising 1422 nucleotides of CEA mRNA. This sequence encodes the 372 COOH-terminal amino acids of CEA followed by 305 nucleotides of 3' untranslated sequence containing a truncated *Alu* repeat. The predicted protein sequence is composed of two repeats comprising 178 amino acids, each with an exceptionally high homology of 67%. Each repeat unit contains four conserved cysteine residues and six to nine putative N-glycosylation sites. CEA mRNA is most strongly expressed in primary colon tumors and, to a lesser extent, in normal colonic tissue. No CEA mRNA is found in HeLa cells and normal human fibroblasts.

One of the most widely used human tumor markers is the carcinoembryonic antigen (CEA). Although this oncodevelopmentally regulated antigen lacks absolute tumor specificity, the CEA concentration in sera of cancer patients is an important parameter in the surveillance of patients in the postoperative phase (1). The presence of CEA on tumor cells has been exploited for immunolocalization of primary tumors and their metastases (2) and for immunotherapy without (3) and with drug targeting (4).

CEA is a heterogenous highly glycosylated protein with a molecular weight of ≈180,000. Highest tissue concentrations are found in adenocarcinomas of the digestive system and in fetal colon. It is encoded by an mRNA of ≈3100 nucleotides and is synthesized as an 80-kDa precursor protein (5). There are some indications that a tumor-specific CEA molecule might exist among the different CEA variants, which is associated only with malignant diseases (6). Recent findings suggest that lung and breast tumors might also contain tissue-specific CEA species (7). In addition to CEA, there are a number of immunologically closely related antigens (e.g., nonspecific cross-reacting antigen or NCA), which vary in size (8) and tissue distribution (9). These cross-reacting antigens are encoded by separate genes, as implied by their very similar but distinct NH_2-terminal amino acid sequences (10).

To define more precisely the various members of the CEA family and to investigate the role of CEA in ontogenesis and in malignancy, we have cloned CEA mRNA. In the present study, we describe the nucleotide sequence of two overlapping CEA cDNA clones (pCEA1 and pCEA2) comprising 1422 nucleotides of the CEA mRNA, the composition of the encoded polypeptide of two large and very closely related repeats, and the expression of CEA mRNA in various normal and tumorous tissues and cell lines.

MATERIALS AND METHODS

Tissues and Cell Lines. After surgical resection, primary human colon adenocarcinomas and nontumorous colonic tissue were separated, frozen, and stored in liquid nitrogen. For some experiments, a human colon carcinoma was used that had been propogated and amplified in athymic mice (11). The colon tumor cell lines SW403 (a gift from J. Shively, Duarte, CA) and LoVo (American Type Culture Collection) were grown in RPMI medium with 10% fetal calf serum and in F-12 medium with 20% fetal calf serum, respectively. The cells were stored in liquid nitrogen until used for RNA isolation.

Construction of a Colon Tumor cDNA Library. Poly(A)$^+$ RNA was isolated from a human colon tumor grown in athymic mice (11) according to a method developed by Fiddes and Goodman (12), as described (13). cDNA synthesis, dC-tailing of double-stranded cDNA, and annealing with *Pst* I linearized and dG-tailed pBR322 (New England Nuclear) were performed essentially as described by Gubler and Hoffman (14). The recombinant plasmids were transfected into competent bacteria of the *Escherichia coli* strain RR1ΔM15 (15) using the protocol of Hanahan (16). About 5000 independent recombinant clones were established on nitrocellulose filters and grown overnight on agar plates in the presence of tetracycline. The colonies were rinsed off the filters with LB and the bacteria were pelleted, resuspended in LB containing 50% (vol/vol) glycerol, and stored at −70°C.

Screening of the cDNA Library. Recombinant clones of the colon tumor cDNA library were grown on nitrocellulose filters, amplified in the presence of chloramphenicol, and screened with synthetic oligonucleotides (provided by J. Shively), essentially as described (17). The sequences of the synthetic oligonucleotides between 42 and 54 nucleotides long (18) had been deduced from partial CEA and NCA amino acid sequences (19) by using a method published by Lathe (20). Colony hybridization with the ^{32}P-labeled oligonucleotides was performed in the presence of 1.08 M NaCl at temperatures 20°C below the melting temperatures calculated for the different oligonucleotides (20). The final stringent wash was in 2× SSPE (1× SSPE = 0.18 M NaCl/10 mM sodium phosphate, pH 7.4/1 mM EDTA)/0.5% NaDodSO₄

Abbreviations: CEA, carcinoembryonic antigen; NCA, nonspecific cross-reacting antigen.

for 15 min at 2°C–4°C below the calculated melting temperatures.

Positive colonies were subjected to a second round of screening; they were hybridized either with ^{32}P-labeled human DNA used as a probe for repetitive sequences (21) or with a 1.4-kilobase (kb) *Eco*RI fragment from the genomic clone λ39.2 of a human NCA gene that had been subcloned in pUC18 [obtained from J. Thompson and J. Shively (18)]. After hybridization, the colony filters were washed under moderately stringent conditions in 2× SSPE/0.5% NaDodSO$_4$ (human repetitive probe) or at high stringency in 0.1× SSPE/0.5% NaDodSO$_4$ at 65°C for 60 min.

RNA Blot Hybridization. Total RNAs or poly(A)-containing RNAs were size-fractionated by electrophoresis on a denaturing methylmercury hydroxide agarose gel. The gel was prepared for transfer of RNA to nitrocellulose according to a procedure described by Alwine *et al.* (22). Hybridization with the cDNA inserts of pCEA1, labeled by random hexanucleotide priming (23), was carried out as described (24). The bands on the autoradiographs were quantitated with an LKB 2202 Ultroscan laser densitometer in connection with an LKB 2200 recording integrator.

DNA Sequencing and Sequence Analysis. Sequencing of cDNAs was performed by the chemical degradation method described by Maxam and Gilbert (25) and by the dideoxy-chain-termination method of Sanger *et al.* (26). For comparison of nucleotide and amino acid sequences, the EMBL nucleotide sequence data bank,* the Protein Identification Resource,† and the Newat 85 protein sequence library obtained from R. R. Doolittle (San Diego, CA) were used. For calculation of the homology of the CEA amino acid repeats, logarithm of odds matrix scores ≥8 and structure genetic matrix scores ≥5 (27) were taken as conservative exchanges.

RESULTS

Isolation of Human CEA cDNA Clones. For isolation of cDNA clones encoding CEA and closely related antigens, we constructed a cDNA library from colon tumor RNA. CEA mRNA comprised 0.25–0.5% of total RNA, as shown by *in vitro* translation and immunoprecipitation of the CEA precursor protein (13). A total of 20,000 cDNA colonies representing 5000 independent clones were screened by hybridization with six different synthetic oligonucleotides (18). Their sequences had been deduced from partial amino acid sequences of CEA and the closely related antigen NCA (19). Of the 120 positive colonies identified, none hybridized with more than one oligonucleotide. We therefore assayed the candidate clones in a second round of selection for the presence of repetitive sequences, because CEA mRNA contains an *Alu*-type repeat (28). Of these candidate clones selected by hybridization with oligonucleotides, 23 hybridized with total human DNA used as a probe for repetitive sequences (21). Seven of the clones positive with the repetitive probe hybridized with oligonucleotide 1, and 13 hybridized with oligonucleotide 2. At this stage of screening, we obtained subcloned fragments of λ39.2, which contains part of a human NCA gene (18). A 1.4-kb *Eco*RI fragment of this genomic clone containing the region complementary to oligonucleotide 2 was chosen for hybridization with the clones preselected by hybridization with the six different oligonucleotides. Thirty positive clones were obtained. As expected, all of these clones had been selected by hybridization with

oligonucleotide 2. Eight of the most strongly hybridizing colonies were colony-purified and their plasmids were isolated. All recombinant plasmids contained identical cDNA inserts when released from the vector by digestion with *Pst* I. They were designated pCEA1. One clone, designated pCEA2, was selected that hybridized with the 1.4-kb *Eco*RI fragment of the NCA clone and with human repetitive sequences. The cDNA of this clone was expected to further extend the CEA cDNA sequence toward the 3' end of the mRNA, assuming the *Alu* sequence to be present in the 3' nontranslated region, as has been reported for other mRNAs (29–31).

Nucleotide Sequence of the Human CEA cDNAs. The extent of overlap of pCEA1 and pCEA2 was determined by restriction endonuclease mapping and sequencing (Fig. 1*B*). Suitable DNA fragments were subcloned and sequenced according to the strategy shown in Fig. 1*B*. The nucleotide sequence is presented in Fig. 1*C*. The sequence can be translated into one open reading frame comprising the 372 COOH-terminal amino acids. Comparison of the deduced amino acid sequence with sequences obtained from CEA and NCA peptides (19) clearly identifies pCEA1 and pCEA2 as CEA rather than NCA cDNA clones (data not shown). The putative COOH-terminal polypeptide of CEA is composed of two very closely related repeats of 178 amino acids having 67% of their amino acids in common (Fig. 2*A*). If one allows for conservative exchanges, the homology increases to 87%. The high degree of homology is also obvious at the nucleotide level. Eighty percent of the nucleotide sequences of the repeats are identical. Due to the presence of the repeating units, oligonucleotide 2, which was used for the initial screening, hybridizes with two different regions in pCEA1 and with one in pCEA2, showing 71% and 74% homology, respectively (Fig. 1*A*). The degree of homology is significantly lower than expected [85% (20)]. This is due to the differences between the amino acid sequence of the peptide from which the nucleotide sequence of oligonucleotide 2 was deduced and the amino acid sequences of the corresponding regions of repeats a and b (Fig. 1*A*). However, this homology was significant enough to yield strong hybridization signals with positive clones (data not shown). The repeat most proximal to the COOH terminus is followed by a unique hydrophobic sequence of 27 amino acids.

The region of CEA that is encoded in pCEA1 and pCEA2 contains 15 putative N-glycosylation sites (Asn-Xaa-$^{Ser}_{Thr}$; Fig. 1*C*), which are clustered around the cysteine residues (Fig. 1*B*). This rather large number of glycosylation sites is to be expected when taking into account the 60% carbohydrate content of mature CEA (33), whose oligosaccharide chains are all linked to asparagine via N-glycosidic bonds (34).

The 3' nontranslated region of CEA mRNA contains part of an *Alu*-type repetitive sequence (Fig. 1 *B* and *C*) as predicted by the observation that two unrelated cDNA clones containing *Alu* repeats hybridize with CEA mRNA (28). Human *Alu* DNA is a head-to-tail dimer of two similar sequences of ≈130 base pairs (bp). Both monomers can be distinguished by the presence of a 31-bp insert in the right repeat (35). By this criterion, only part of an *Alu* repeat is present in CEA mRNA, with the left monomer being truncated at the 5' end.

Comparison of the nucleotide sequence and the deduced amino acid sequence of pCEA1 and pCEA2 with sequence data banks did not reveal any significant homology to other sequences.

RNA Blot Analysis of CEA mRNA Expression. The cDNA fragments of pCEA1 were used for hybridization with blots of size-fractionated total RNA isolated from normal and tumorous cells and tissues. As shown in Fig. 3, the cDNA of plasmid pCEA1 hybridizes to two mRNA species of 3500 and

---

*European Molecular Biology Laboratory (1985) EMBL Nucleotide Sequence Data Library (Eur. Mol. Biol. Lab., Heidelberg, F.R.G.), Release 7.
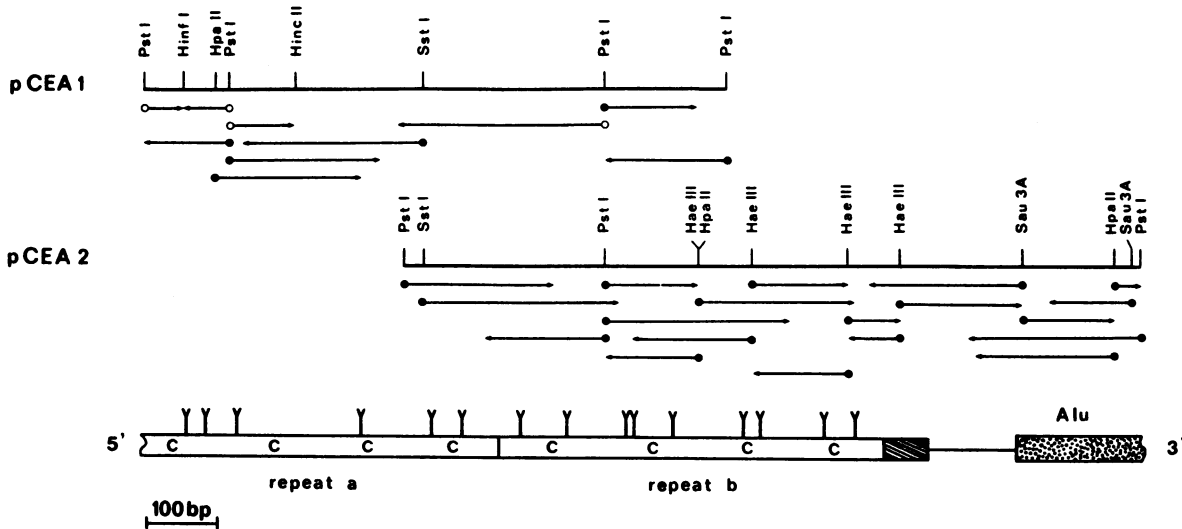
†Protein Identification Resource (1986) Protein Sequence Database (Natl. Biomed. Res. Found., Washington, DC), Release 8.0.

A

```
CEA peptide                    N   D   A   G   S   Y   E   C   E   I   Q   N   P   A
derived mRNA sequence   5'- AAU GAU GCU GGC UCC UAU GAA UGU GAA AUU CAG AAC CCA GCC -3'
oligonucleotide 2       3'- TTA CTA CGA CCG AGG ATA CTT ACA CTT TAA GTC TTG GGT CGG -5'

repeat a       pos. 167   ... ... .AT ..T G.. ... ..C ... .C. ..G ... ... CT. AAT
repeat b       pos. 701   ... ..G ..T T.T C.. ... .A. ... .C. ..G ... ... A.. .AC
```

B

C

```
   1 CAACTCCAACCCCGTGGAGGATGAGGATGCTGTAGCCTTAACCTGTGAACCTGAGATTCAGAACACAACCTACCTGTGGTGGGTAAATAATCAGAGCCTCCCGGTCAGTCCCAGGCTGCA
   1 AsnSerAsnProValGluAspGluAspAlaValAlaLeuThrCysGluProGluIleGln[AsnThrThr]TyrLeuTrpTrpValAsn[AsnGlnSer]LeuProValSerProArgLeuGln

 121 GCTGTCCAATGACAACAGGACCCTCACTCTACTCAGTGTCACAAGGAATGATGTAGGACCCTATGAGTGTGGAATCCAGAACGAATTAAGTGTTGACCACAGCGACCCAGTCATCCTGAA
  41 LeuSerAsnAsp[AsnArgThr]LeuThrLeuLeuSerValThrArgAsnAspValGlyProTyrGluCysGlyIleGlnAsnGluLeuSerValAspHisSerAspProValIleLeuAsn

 241 TGTCCTCTATGGCCCAGACGACCCCACCATTTCCCCCTCATACACCTATTACCGTCCAGGGGTGAACCTCAGCCTCTCCTGCCATGCAGCCTCTAACCCACCTGCACAGTATTCTTGGCT
  81 ValLeuTyrGlyProAspAspProThrIleSerProSerTyrThrTyrTyrArgProGlyVal[AsnLeuSer]LeuSerCysHisAlaAlaSerAsnProProAlaGlnTyrSerTrpLeu

 361 GATTGATGGGAACATCCAGCAACACACACAAGAGCTCTTTATCTCCAACATCACTGAGAAGAACAGCGGACTCTATACCTGCCAGGCCAATAACTCAGCCAGTGGCCACAGCAGGACTAC
 121 IleAspGlyAsnIleGlnGlnHisThrGlnGluLeuPheIleSer[AsnIleThr]GluLysAsnSerGlyLeuTyrThrCysGlnAla[AsnAsnSer]AlaSerGlyHisSerArgThrThr

                     ┌─────── repeat b
 481 AGTCAAGACAATCACAGTCTCTGCGGAGCTGCCCAAGCCCTCCATCTCCAGCAACAACTCCAAACCCGTGGAGGACAAGGATGCTGTGGCCTTCACCTGTGAACCTGAGGCTCAGAACAC
 161 ValLysThrIleThrValSerAlaGluLeuProLysProSerIleSerSer[AsnAsnSer]LysProValGluAspLysAspAlaValAlaPheThrCysGluProGluAlaGln[AsnThr]

 601 AACCTACCTGTGGTGGGTAAATGGTCAGAGCCTCCCAGTCAGTCCCAGGCTGCAGCTGTCCAATGGCAACAGGACCCTCACTCTATTCAATGTCACAAGAAATGACGCAAGAGCCTATGT
 201 [Thr]TyrLeuTrpTrpValAsnGlyGlnSerLeuProValSerProArgLeuGlnLeuSerAsnGly[AsnArgThr]LeuThrLeuPhe[AsnValThr]ArgAsnAspAlaArgAlaTyrVal

 721 ATGTGGAATCCAGAACTCAGTGAGTGCAAACCGCAGTGACCCAGTCACCCTGGATGTCCTCTATGGGCCGGACACCCCCATCATTTCCCCCCCAGACTCGTCTTACCTTTCGGGAGCGAA
 241 CysGlyIleGlnAsnSerValSerAla[AsnArgSer]AspProValThrLeuAspValLeuTyrGlyProAspThrProIleIleSerProProAspSerSerTyrLeuSerGlyAlaAsn

 841 CCTCAACCTCTCCTGCCACTCGGCCTCTAACCCATCCCCGCAGTATTCTTGGCGTATCAATGGGATACCGCAGCAACACACACAAGTTCTCTTTATCGCCAAAATCACGCCAAATAATAA
 281 Leu[AsnLeuSer]CysHisSerAlaSer[AsnProSer]ProGlnTyrSerTrpArgIleAsnGlyIleProGlnGlnHisThrGlnValLeuPheIleAlaLysIleThrProAsnAsn[Asn]
                                                                                 .end repeat b ───────┐
 961 CGGGACCTATGCCTGTTTTGTCTCTAACTTGGCTACTGGCCGCAATAATTCCATAGTCAAGAGCATCACAGTCTCTGCATCTGGAACTTCTCCTGGTCTCTCAGCTGGGGCCACTGTCGG
 321 [GlyThr]TyrAlaCysPheValSerAsnLeuAlaThrGlyArg[AsnAsnSer]IleValLysSerIleThrValSerAlaSerGlyThrSerProGlyLeuSerAlaGlyAlaThrValGly

1081 CATCATGATTGGAGTGCTGGTTGGGGGTTGCTCTGATATAGCAGCCCTGGTGTAGTTTCTTCATTTCAGGAAGACTGACAGTTGTTTTGCTTCTTCCTTAAAGCATTTGCAACAGCTACAG
 361 IleMetIleGlyValLeuValGlyValAlaLeuIle***

1201 TCTAAAATTGCTTCTTTACCAAGGATATTTACAGAAAAGACTCTGACCAGAGATCGAGACCATCCTAGCCAACATCGTGAAACCCCATCTCTACTAAAAATACAAAAATGAGCTGGGCTT

1321 GGTGGCGCGCACCTGTAGTCCCAGTTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGGAGGTGGAGATTGCAGTGAGCCCAGATCGCACCACTGC
```

FIG. 1.   (A) Amino acid sequence (designated by the single-letter code) of a CEA peptide and the deduced nucleotide sequence of oligonucleotide 2, which is complementary to CEA mRNA (19). The nucleotide sequence of oligonucleotide 2 is compared to the homologous sequences of repeats a and b of CEA cDNA. Dots denote identical bases. Numbers indicate positions of the CEA cDNA region that is homologous to oligonucleotide 2. (B) Structure of the overlapping CEA cDNA clones pCEA1 and pCEA2. A restriction map of various endonuclease sites is shown. For the restriction endonucleases *Hae* III, *Hpa* II, and *Hin*fI, only the sites relevant for sequencing are included. The *Pst* I sites flanking the cDNA fragments are artificially created by the cloning procedure. Direction and extent of sequence analysis are indicated by arrows. Sanger (●) and Maxam–Gilbert (○) sequencing protocols were used. Below the sequencing strategy, several features of the CEA mRNA are summarized graphically. Open boxes symbolize repeats a and b of the coding region. The positions of the cysteine residues are indicated by C, and the locations of the putative N-glycosylation sites are indicated by Υ. Hatched box depicts the unique hydrophobic COOH terminus of CEA. The truncated *Alu* sequence in the 3' noncoding region is shown as a stippled box. (C) Nucleotide sequence and predicted amino acid sequence of the human CEA cDNA clones pCEA1 and pCEA2. The putative N-glycosylation sites are boxed. The *Alu* sequence in the 3' nontranslated region is underlined.
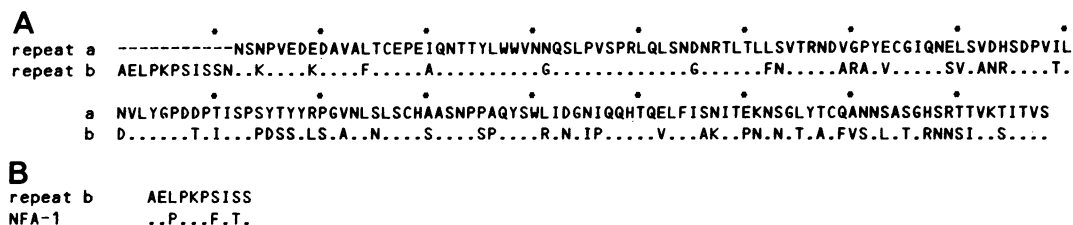
A

```
repeat a    -----------NSNPVEDEDAVALTCEPEIQNTTYLWWVNNQSLPVSPRLQLSNDNRTLTLLSVTRNDVGPYECGIQNELSVDHSDPVIL
repeat b    AELPKPSISSN..K....K....F.....A...........G..............G......FN.....ARA.V.....SV.ANR....T.

        a   NVLYGPDDPTISPSYTYYRPGVNLSLSCHAASNPPAQYSWLIDGNIQQHTQELFISNITEKNSGLYTCQANNSASGHSRTTVKTITVS
        b   D......T.I...PDSS.LS.A..N....S....SP....R.N.IP.....V...AK..PN.N.T.A.FVS.L.T.RNNSI..S....
```

B

```
repeat b    AELPKPSISS
NFA-1       ..P...F.T.
```

FIG. 2. (A) Comparison of repeats a and b of the predicted amino acid sequence encoded by pCEA1 and pCEA2. Identical residues are indicated by dots. Dashes denote region of repeat a that is missing in pCEA1. The presence of a sequence similar to the homologous region of repeat b is inferred from amino acid sequence data (R. Paxton, personal communication). (B) Comparison of the NH₂-terminal regions of repeat b of CEA with the NH₂ terminus of NFA-1 (32). Dots denote identical residues. Amino acids are identified by the single-letter code.

3000 nucleotides, the smaller species being ≈5-fold more abundant than the larger species. CEA mRNA is most strongly expressed in human primary colon tumors tested either directly after resection or after passage through athymic nude mice (Fig. 3A). CEA mRNA is found to a lesser extent in the colon tumor cell lines LoVo and SW403 and in nontumorous colonic tissue from colon tumor patients (Fig. 3B). No hybridization was detected with poly(A)-containing RNA isolated from HeLa cells and normal human fibroblasts (data not shown).

DISCUSSION

Here we describe the isolation and characterization of two overlapping CEA cDNA clones. The most striking feature of the CEA mRNA is the repeated nature of the coding region so far identified. The open reading frame encoded in pCEA1 and pCEA2 consists of two very closely related repeats containing 178 amino acids, each without any deletion or insertion. Both repeats share 67% of their amino acid sequence. This exceptionally high conservation of the amino acid sequence of CEA is much more pronounced than in other vertebrate proteins composed of repeated domains. Generally, the homology of internal repeats of vertebrate proteins is only 20–40% [e.g., human low density lipoprotein receptor (29), human α₁ B-glycoprotein (36), and human

FIG. 3. Analysis of CEA mRNA in normal and tumorous colon tissues and colon tumor cell lines. Two micrograms of poly(A)⁺ RNA (A) or 10 μg of total RNA (B) was separated on a 1% agarose/methylmercury hydroxide gel by electrophoresis. The RNA was transferred to nitrocellulose, and the filters were hybridized with ³²P-labeled cDNA inserts of pCEA1. The RNAs had been isolated from primary colon tumors (A, lanes 2–4) of various patients (indicated by the letters above lanes), from a primary colon tumor propagated in nude mice (A, lane 1; B, lane 3), from the colon tumor cell lines SW403 (B, lane 1) and LoVo (B, lane 2), or from normal colonic tissue from a colon tumor patient (B, lane 4). The high molecular weight bands in lanes 2 and 4 in B are due to unspecific binding of the probe to DNA present in these RNA preparations. The numbers on the left indicate the size (in kb) and positions of marker DNAs and RNAs.

fibronectin repeats I and III (37)]. To our knowledge, the highest homology between internal repeats observed to date exists between the two type II repeats in human fibronectin [50% (37)] and in the two repeats within human transferrin [45% (38)]. Normally, the length of internal repeats is rather short, varying between 40 and 95 amino acids (39). The human transferrin and albumin, however, are composed of repeated domains of 360 and 195 amino acids, respectively (39). With a repeat length of 178 amino acids, CEA belongs to this class of proteins with internal repeats.

The high degree of homology between the CEA repeats is even more pronounced at the nucleotide level. The strong conservation of both amino acid and nucleotide sequences implies that the CEA gene evolved rather recently in evolution by duplication of a primordial gene. Even more recently, the various very closely related members of the CEA gene family, comprising about one dozen genes (18), must have arisen by further gene duplication or exon shuffling (40). This model was already proposed by Grunert *et al.* (8), who reported that CEA and related antigens are probably composed of various numbers of similar domains as evidenced by the almost identical fingerprints of two CEA and three NCA species having molecular weights between 50,000 and 180,000. The presence of closely related repeats in CEA explains the finding that some monoclonal anti-CEA antibodies bind to two or more epitopes on the same CEA molecule (ref. 41; K. Schwarz, personal communication). At present, the total number of repeats contained within CEA is not known. However, the presence of probably three 178-amino acid repeats can be expected, taking into account the size of the unglycosylated CEA precursor [≈700 amino acids (5)], amino acid sequence data (19), and the length of the unique NH₂-terminal sequence of a closely related NCA gene [107 amino acids (18)]. It is interesting to note that the repeats identified by CEA cDNA sequencing might be identical with protease-resistant domains of CEA or related antigens. The NH₂-terminal amino acid sequence of the normal fecal antigen 1 (NFA-1), a protein closely related to CEA (32), shares 70% homology with the region of the NH₂-terminal sequence of the repeat unit of CEA (Fig. 2B). NFA-1 has a molecular weight of 20,000–30,000 and might be a naturally occurring degradation product of a larger CEA-related antigen (32).

In each repeat, the positions of four cysteine residues are conserved and nearly equidistant (39–47 amino acids between the residues), implying a very similar three-dimensional structure of both protein domains with regularly spaced folds. In contrast to the strongly conserved positions of the cysteine residues, the number and location of the N-glycosylation sites are poorly conserved, with only two sites being at homologous positions within the repeats. This lack of conservation of homologous N-glycosylation sites has also been reported for immunoglobulins and other plasma glycoproteins composed of repeating domains (36).

Although most of the CEA (80%) is associated with the membrane fraction (42), it is still disputed whether CEA is an
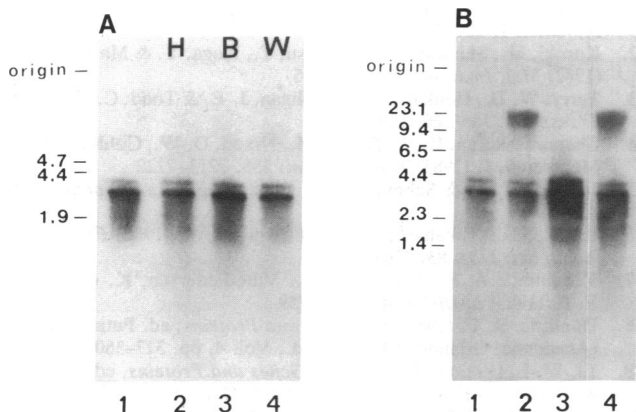
integral membrane protein or not. By ultrastructural studies, CEA has been located primarily in the intestinal glycocalyx (43). The presence of a hydrophobic COOH-terminal region of the putative CEA polypeptide suggests that the CEA encoded by pCEA1 and pCEA2 is anchored in the plasma membrane with an extracellular orientation of the NH_2 terminus. The latter is inferred by the observation that CEA can be labeled by the lactoperoxidase method on the surface of intact colon tumor cells (R. Bhardwaj, personal communication).

There is one methionine residue in the unique COOH-terminal region. Until now, the absence of methionine from CEA, as demonstrated, e.g., by the resistance of CEA to cyanogen bromide cleavage (44), was used as one of the identification criteria of CEA (45). Our data predict that cyanogen bromide treatment of CEA would release a 10-amino acid fragment from the COOH terminus of CEA. This small peptide could be easily missed by conventional protein analysis.

The 3' untranslated region of CEA mRNA contains a truncated *Alu* repeat sequence. With >300,000 copies present per haploid genome, *Alu* sequences are the most abundant family of human repetitive DNA sequences (35). The dispersal of *Alu* elements throughout gene regions has been reported for a number of genes (35). However, to date only a few mRNAs have been found to carry *Alu* sequences. The mRNAs for the human low density lipoprotein receptor (29), the human p53 cellular tumor antigen (30), and the murine class I histocompatibility antigen of the D region (31) contain one or more *Alu* monomers or dimers exclusively in the 3' untranslated region. At present, it is not known whether repetitive sequences in mRNAs or in the genome have a function (35).

CEA mRNA is expressed in a tissue-specific manner. We observed no expression in HeLa tumor cells or normal human fibroblasts, whereas colonic tissues and cell lines contain CEA mRNA. The highest levels of CEA mRNA are found in primary colon and rectum carcinomas. Adaptation to growth in athymic nude mice does not change the CEA gene expression significantly. However, establishment of primary tumors as cell lines strongly reduces the CEA mRNA content. The relatively strong expression of CEA mRNA in normal colonic tissue is somewhat surprising, although CEA-like antigens have been identified in normal colonic mucosa (46) or in colon lavages of healthy individuals (47). Whether this relatively high level of CEA mRNA observed in normal colonic tissue of tumor patients is due to inflammation or to the presence of tumor cells not visible macroscopically in areas surrounding the tumor remains to be established.

1. Gold, P. & Shuster, J. (1981) *Cancer Res.* **40**, 2973–2976.
2. Goldenberg, D. M., DeLand, F., Kim, E., Bennett, S., Primus, F. J., van Nagell, J. R., Jr., Estes, N., DeSimone, P. & Rayburn, P. (1978) *N. Engl. J. Med.* **298**, 1384–1388.
3. Goldenberg, D. M., Gaffar, S. A., Bennett, S. J. & Beach, J. L. (1981) *Cancer Res.* **41**, 4354–4360.
4. Pagé, M., Delorme, F., Lafontaine, F. & Dumas, L. (1984) *Semin. Oncol.* **11**, 56–58.
5. Zimmermann, W., Thompson, J., Grunert, F., Luckenbach, G. A., Friedrich, R. & von Kleist, S. (1984) in *Genes and Antigens in Cancer Cells: The Monoclonal Antibody Approach*, eds. Riethmüller, G., Koprowski, H., von Kleist, S. & Munk, K. (Karger, Basel), Vol. 19, pp. 64–74.
6. Jothy, S., Brazinsky, S. A., Chin-A-Loy, M., Haggarty, A., Krantz, M. J., Cheung, M. & Fuks, A. (1986) *Lab. Invest.* **54**, 108–117.
7. Grunert, F., Luckenbach, G. A., Haderlie, B., Schwarz, K. & von Kleist, S. (1983) *Ann. N.Y. Acad. Sci.* **417**, 75–85.
8. Grunert, F., AbuHarfeil, N., Schwarz, K. & von Kleist, S. (1985) *Int. J. Cancer* **36**, 357–362.
9. Buchegger, F., Schreyer, M., Carrel, S. & Mach, J. P. (1984) *Int. J. Cancer* **33**, 643–639.
10. Kessler, M. H., Shively, J. E., Pritchard, D. G. & Todd, C. W. (1978) *Cancer Res.* **38**, 1041–1048.
11. Fiebig, H. H. & Löhr, G. W. (1984) *Med. Welt* **35**, 52–58.
12. Fiddes, J. C. & Goodman, H. M. (1979) *Nature (London)* **281**, 351–356.
13. Zimmermann, W., Friedrich, R., Grunert, F., Luckenbach, G. A., Thompson, J. & von Kleist, S. (1983) *Ann. N.Y. Acad. Sci.* **417**, 21–30.
14. Gubler, U. & Hoffman, B. J. (1983) *Gene* **25**, 263–269.
15. Rüther, U., Koenen, M., Otto, K. & Müller-Hill, B. (1981) *Nucleic Acids Res.* **16**, 4087–4098.
16. Hanahan, D. (1983) *J. Mol. Biol.* **166**, 557–580.
17. Whitehead, A. S., Goldberger, G., Woods, D. E., Markham, A. F. & Colten, H. R. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 5387–5391.
18. Thompson, J., Pande, H., Paxton, R., Shively, L., Padma, A., Simmer, R., Todd, C. W., Riggs, A. D. & Shively, J. E. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 2965–2969.
19. Paxton, R., Mooser, G., Pande, H., Lee, T. D. & Shivley, J. E. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 920–924.
20. Lathe, R. (1985) *J. Mol. Biol.* **183**, 1–12.
21. Wu, B., Hunt, C. & Morimoto, R. (1985) *Mol. Cell. Biol.* **5**, 330–341.
22. Alwine, J., Kamp, D. J. & Stark, G. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5350–5354.
23. Feinberg, A. P. & Vogelstein, B. (1983) *Anal. Biochem.* **132**, 6–13.
24. Thomas, P. S. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5201–5205.
25. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
26. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
27. Feng, D. F., Johnson, M. S. & Doolittle, R. F. (1985) *J. Mol. Evol.* **21**, 112–125.
28. Zimmermann, W., Weiss, M., Friedrich, R. & von Kleist, S. (1985) *Tumor Biol.* **6**, 323 (abstr.).
29. Yamamoto, T., Davis, C. G., Brown, M. S., Schneider, W. J., Casey, M. L., Goldstein, J. L. & Russel, D. W. (1984) *Cell* **39**, 27–38.
30. Matlashewski, G., Lamb, P., Pim, D., Peacock, J., Crawford, L. & Benchimol, S. (1984) *EMBO J.* **3**, 3257–3262.
31. Hood, L., Steinmetz, M. & Malissen, B. (1983) *Annu. Rev. Immunol.* **1**, 529–568.
32. Kuroki, M., Shinoda, T., Takayasu, T., Koga, Y. & Matsuoka, Y. (1982) *Mol. Immunol.* **19**, 399–406.
33. Terry, W. D., Henkart, P. A., Coligan, J. E. & Todd, C. W. (1974) *Transplant. Rev.* **20**, 100–129.
34. Chandrasekaran, E. V., Davila, M., Nixon, D. W., Goldfarb, M. & Mendicino, J. (1983) *J. Biol. Chem.* **258**, 7213–7222.
35. Jelinek, W. R. & Schmid, C. W. (1982) *Annu. Rev. Biochem.* **51**, 813–844.
36. Ishioka, N., Takahashi, N. & Putnam, F. W. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 2363–2367.
37. Kornblihtt, A. R., Umezawa, K., Vibe-Pedersen, K. & Baralle, F. E. (1985) *EMBO J.* **4**, 1755–1759.
38. Doolittle, R. F. (1984) in *The Plasma Proteins*, ed. Putnam, S. W. (Academic, Orlando, FL), 2nd Ed., Vol. 4, pp. 317–360.
39. Li, W.-L. (1983) in *Evolution of Genes and Proteins*, eds. Nei, M. & Koehn, R. K. (Sinauer, Sunderland, MA), pp. 14–37.
40. Gilbert, W. (1978) *Nature (London)* **271**, 501.
41. Hedin, A., Hammarström, S. & Larsson, A. (1982) *Mol. Immunol.* **19**, 1641–1648.
42. Shi, Z. R., Tsao, D. & Kim, Y. S. (1983) *Cancer Res.* **43**, 4045–4049.
43. Gold, P., Krupey, J. & Ansari, H. (1970) *J. Natl. Cancer Inst.* **45**, 219–225.
44. AbuHarfeil, N., Grunert, F. & von Kleist, S. (1984) *Tumor Biol.* **5**, 339–350.
45. Plow, E. F. & Edgington, T. S. (1979) in *Immunodiagnosis of Cancer*, eds. Herberman, R. B. & McIntire, K. R. (Dekker, New York), Part I, pp. 181–239.
46. Fritsche, R. & Mach, J.-P. (1978) *Immunochemistry* **14**, 119–127.
47. Shively, J. E., Todd, C. W., Go, V. L. W. & Egan, M. L. (1978) *Cancer Res.* **38**, 503–505.