MOLECULAR AND CELLULAR BIOLOGY, Sept. 1987, p. 3221–3230
0270-7306/87/093221-10$02.00/0

Vol. 7, No. 9

Isolation and Characterization of Full-Length Functional cDNA Clones for Human Carcinoembryonic Antigen

NICOLE BEAUCHEMIN,[1] SARITA BENCHIMOL,[1] DENIS COURNOYER,[1] ABRAHAM FUKS,[2] AND CLIFFORD P. STANNERS[1,2]*

Department of Biochemistry[1] and McGill Cancer Centre,[2] McGill University, Montreal, Quebec H3G 1Y6, Canada

Received 14 April 1987/Accepted 11 June 1987

Carcinoembryonic antigen (CEA) expression is perhaps the most prevalent of phenotypic changes observed in human cancer cells. The molecular genetic basis of this phenomenon, however, is completely unknown. Twenty-seven CEA cDNA clones were isolated from a human colon adenocarcinoma cell line. Most of these clones are full length and consist of a number (usually three) of surprisingly similar long (534 base pairs) repeats between a 5′ end of 520 base pairs and a 3′ end with three different termination points. The predicted translation product of these clones consists of a processed signal sequence of 34 amino acids, an amino-terminal sequence of 107 amino acids, which includes the known terminal amino acid sequence of CEA, three repeated domains of 178 amino acids each, and a membrane-anchoring domain of 27 amino acids, giving a total of 702 amino acids and a molecular weight of 72,813 for the mature protein. The repeated domains have conserved features, including the first 67 amino acids at their N termini and the presence of four cysteine residues. Comparisons with the amino acid sequences of other proteins reveals homology of the repeats with various members of the immunoglobulin supergene family, particularly the human T-cell receptor γ chain. CEA cDNA clones in the SP-65 vector were shown to produce transcripts in vitro which could be translated in vitro to yield a protein of molecular weight 73,000 which in turn could be precipitated with CEA-specific antibodies. CEA cDNA clones were also inserted into an animal cell expression vector and introduced by transfection into mammalian cell lines. These transfectants produced a CEA-immunoprecipitable glycoprotein which could be visualized by immunofluorescence on the cell surface.

Carcinoembryonic antigen (CEA), a large cell membrane glycoprotein of molecular weight about 180,000 (40), is produced in a high proportion of human tumors arising at the most common sites including colon, breast, and lung (37). CEA is also produced during human embryogenesis, while a related but distinctly different series of antigens termed CEA-cross-reactive antigens can be produced by a variety of adult normal cells (17). The consistent presence of CEA in tumors has led to its wide use as a clinical assay for prognosis and management of colon carcinomas. The normal function, role in carcinogenesis, and molecular basis of production of CEA in tumors, however, are completely unknown. As a first step in the solution of these questions, we have initiated a study of the molecular genetics of CEA, beginning with the isolation of a family of CEA cDNA clones.

We now report the molecular cloning, nucleotide sequence, and structural analysis of a number of functional full-length CEA cDNA clones. These clones reveal a basic structure of common 5′ and 3′ ends embracing a number (usually three) of strikingly similar relatively long repeating units. The repeats include amino acid sequences which show significant homology with several members of the rat immunoglobulin family and, in particular, with the variable domain of the human T-cell receptor γ chain. Several of the cDNA clones have been shown to produce CEA both by translation of their transcripts generated in vitro and by the demonstration of CEA production in cells transfected with them when inserted in expression vectors.

While this work was in progress, the isolation and nucleotide sequences of partial CEA cDNA clones were reported by Zimmerman et al. (48) and Oikawa et al. (28) and those of a portion of a genomic clone of normal cross-reacting antigen (NCA), the most common normal counterpart of CEA, were reported by Thompson et al. (41).

MATERIALS AND METHODS

Cell culture. Cells of human colonic adenocarcinoma lines LS174T and LS180 (32) and their subclones, of human embryo fibroblasts, and of the CHO line LR-73 (31) were cultured at 37°C in monolayer in α-minimal essential medium (38) supplemented with 10% fetal bovine serum.

Purification of RNA and Northern blot (RNA blot) analysis. Total RNA was isolated by the guanidium isothiocyanate procedure of Chirgwin et al. (4). Poly(A)$^+$ RNA was purified by two successive passes through oligo(dT)-cellulose by the protocol of Aviv and Leder (1). Samples of total and poly(A)$^+$ RNA were electrophoresed on 1.1 M formaldehyde–1.5% agarose gels (23) and transferred to nitrocellulose filters. Detection of bands with random primer ^{32}P-labeled cDNA probes (11) was done by hybridization (23) for 18 h at 42°C in 5× SSPE (1× SSPE is 0.18 M NaCl plus 10 mM $NaPO_4$ [pH 7.7] plus 1 mM EDTA)–1× Denhardt solution–50% formamide–150 μg of heat-denatured salmon testis DNA per ml–10% dextran sulfate–10^6 cpm of radioactive probe per ml. Filters were washed twice in SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate) for 15 min at 22°C and twice in 0.1× SSC–0.1% sodium dodecyl sulfate at 55°C.

cDNA library and isolation of CEA cDNA clones. cDNA was generated from LS180 poly(A)$^+$ RNA by a modification of the method of Gubler and Hoffman (15). *Eco*RI linkers were ligated to double-stranded cDNA, which was then size selected (≥2 kilobases) after separation on a Bio-Gel A50M column (Bio-Rad Laboratories, Richmond, Calif.), ligated to *Eco*RI-digested λgt10 DNA, and packaged with Gigapack

* Corresponding author.

extracts (Stratagene) to yield infectious virus (23). λgt10 recombinant bacteriophages were plated on *E. coli* C600 (*hsdR hsdM*$^+$ *supE thr leu thi lacY1 tonA21*) or C600 *hflA150* (18). A total of 5×10^5 independent clones (1/10 of the entire library) were screened by duplicate plaque hybridization by using two ^{32}P-labeled oligonucleotide probes (24) (described in Results) applied to nitrocellulose filter images of 529-cm^2 plates with 5×10^4 plaques each. The plaques were amplified on the filters for 6 h at 37°C before probing was done (23). For the 54-mer probe, hybridization was carried out as above at 37°C with 30% (vol/vol) formamide and 10^6 cpm of 5′-end-labeled probe per ml. For the 16-fold redundant 17-mer, hybridization was carried out at 30°C without formamide but including 2 mM sodium PP_i. Filters in both cases were washed in 3 M tetramethylammonium chloride solution at 37°C (46).

To reduce the probability of recombination between repeated nucleotide sequences in CEA cDNA clones, a phenomenon which was observed to artifactually increase or decrease the number of repeats in the clones during propagation, these clones were plaque purified in some cases by using the recombination-deficient strain *E. coli* D1319 (*recD 1014 hsdR2 zjj-202*::Tn*10 recA*::*cam supF58 trp89*::Tn*5*) (42).

5′ end mapping of CEA mRNA. A primer extension reaction was performed by first hybridizing 10 ng of a 5′-end ^{32}P-labeled 21-mer (complementary to the signal sequence) at 55°C with 30 μg of total RNA from LS180 cells in 10 μl of 10 mM PIPES [piperazine-*N*,*N*′-bis(2-ethanesulfonic acid)] buffer (pH 6.4)–0.4 M NaCl. The primer was then extended in 50 mM Tris hydrochloride (pH 8.3)–500 μM deoxynucleoside triphosphates–6 mM $MgCl_2$–10 mM dithiothreitol–100 U of RNasin/ml–350 U of avian myeloblastosis virus reverse transcriptase (Life Sciences, Inc., St. Petersburg, Fla.) per ml in a total volume of 100 μl at 42°C for 1 h; 5 μl of 0.5 M EDTA was added to stop the reaction. The cDNA was phenol extracted, precipitated in ethanol, suspended in 5 μl of 1× Tris-borate-EDTA buffer, and analyzed by electrophoresis on an 8 M urea–7.5% polyacrylamide sequencing gel. A control experiment was carried out in which yeast tRNA replaced the LS180 RNA; no band was obtained (data not shown). To localize the end of the cDNA, we performed a sequencing reaction on the 5′ end of a CEA cDNA clone subcloned in M13 by using the same 5′-end ^{32}P-labeled primer.

DNA sequence determination. DNA fragments to be sequenced were inserted into M13mp18 and mp19 bacteriophage, and single-stranded DNA was sequenced by the dideoxy method of Sanger et al. (33). The entire nucleotide sequence of the cDNA was determined on both strands from three independent λ phage clones.

Labeling and immunoprecipitation of CEA. The *Eco*RI inserts of various CEA cDNA clones were inserted into the *Eco*RI site of the animal cell expression vector p91023B (19, 45). These were introduced into the CHO LR-73 cell line by the calcium phosphate procedure (14). Cultures of the transfectants were labeled with [^{3}H]leucine (155 Ci/mmol) at 100 μCi/ml for 2 h at 37°C in growth medium lacking leucine. Cells were removed from the plastic culture flasks with isotonic phosphate-buffered saline solution containing 17 mM sodium citrate, centrifuged, and washed with phosphate-buffered saline. The cell pellet was solubilized and centrifuged essentially as described by Shore et al. (36). One-half of the supernatant was reacted with a 1/100 dilution of polyclonal rabbit anti-CEA antiserum raised against purified CEA from human colonic tumor metastases (37), while the other half was treated with the same dilution of normal rabbit serum. Immunoprecipitates obtained with protein A-Sepharose (36) were subjected to electrophoresis on 10% polyacrylamide gels containing 0.4% sodium dodecyl sulfate (22). The gel was dried on filter paper and exposed to X-ray film for 3 days.

CEA assay. Cells were disrupted by three 10-s sonication bursts by using an immersion probe. The sonicate was assayed for CEA with the CEA double monoclonal antibody clinical kit (Abbott CEA-EIA Monoclonal; Abbott Laboratories, North Chicago, Ill.). This assay is highly specific and sensitive, since it involves binding to one specific monoclonal antibody on polystyrene beads followed by detection of the bound CEA with a second specific monoclonal antibody. Internal standards allowed calculation of the amount of CEA, which was normalized to the amount of protein in the sonic extracts as measured by the Bio-Rad protein assay.

Sequence analysis. Nucleotide and amino acid sequences were analyzed by using the programs of Deverreux et al. (7) and the ALIGN program of Dayhoff et al. (5). Searches of the National Biomedical Research Foundation Protein Database (6) were performed by using the word search program FastP as described by Wilbur and Lipman (43).

RESULTS

Cloning and characterization of the CEA cDNA family. Our strategy required the development of a series of closely related clonal cell lines with widely varying levels of CEA production to validate oligonucleotide probes representative of known portions of the CEA amino acid sequence. These would also be used to provide a source of RNA enriched in CEA mRNA for preparation of a cDNA library.

When large numbers of individual cell clones randomly picked from the CEA-producing human colon carcinoma cell lines LS180 and LS174T were grown into mass culture and tested for cell-associated CEA by a sensitive double monoclonal antibody-based test, a surprisingly wide variation in CEA levels (about 10^4-fold) was observed (S. Benchimol, L. Bastien, and C. P. Stanners, manuscript in preparation). The parent line, LS180, showed a cell-associated CEA level of about 1,000 ng/mg, or about 0.2% of the total cellular protein, allowing for the equal amount of CEA exported into the medium; for a low-producing clone, clone 86/8, the value was ≤0.00002%.

Total and poly(A)$^+$ RNA preparations from LS180 and clone 86/8 were subjected to Northern analysis with CEA-specific oligonucleotides as probes. Of the probes tested, two gave the expected results, i.e., bands at about 3 kilobases for LS180 mRNA and none for clone 86/8 mRNA (data not shown). The successful probes were a 54-mer, representing a guess of the most likely codons corresponding to the amino acid sequence of one internal fragment of CEA, and a 16-fold degenerate 17-mer, corresponding to an amino acid sequence of another internal fragment (R. J. Paxton, R. L. Simmer, and J. E. Shively, Abstr. Int. Soc. Oncodev. Biol. Med. XIII, abstr. no. A14, p. 47, 1985).

A cDNA library of 5×10^6 independent clones was prepared with size-selected cDNA derived from the total cellular poly(A)$^+$ RNA of LS180 which was inserted into the λgt10 vector. A total of 5×10^5 of these clones were screened with each of the above probes and yielded 57 double positives. Most of these were plaque purified and subjected to further analysis by restriction mapping and identification of restriction fragments which could hybridize with the probes. The restriction map showing the basic

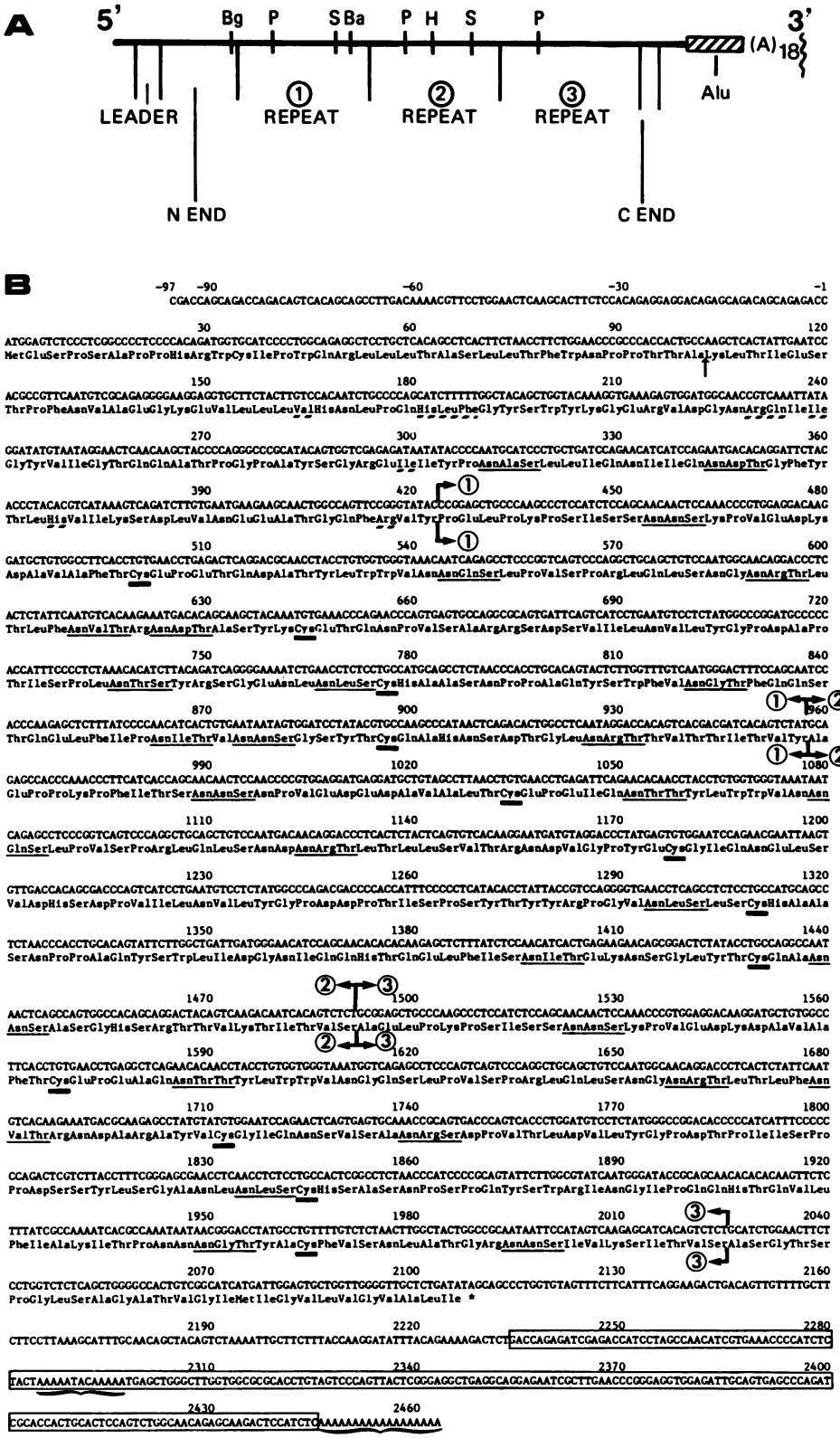

FIG. 1. Structure of CEA cDNA. (A) Restriction map of a typical three-repeat CEA cDNA clone showing the positions of the proposed 5′-terminal leader sequence, the N-terminal end of the mature protein, three very similar repeats, a small C-terminal end of the protein, and an Alu repetitive sequence with a short poly(A) tail at the 3′ terminus. Abbreviations: Bg, *Bgl*II; P, *Pst*I; S, *Sst*I; Ba, *Bam*HI; H, *Hin*cII. (B) Nucleotide sequence and predicted amino acid sequence of the three-repeat CEA cDNA. The nucleotide sequence is listed 5′ to 3′, left to right. Nucleotide positions are shown in overlined numbering; amino acid positions are shown in side numbering. The NH_2-terminal Lys residue of the mature protein is indicated by the vertical arrow. The ends of the three 534-bp (178 amino acid) repeats are indicated by circled numbers. The Alu repetitive sequence is boxed. Dashed underlined amino acid positions indicate differences with NCA-95 in the N-terminal portion of the molecule. Underlined amino acids represent probable N-linked glycosylation sites, and heavily underlined amino acids represent cysteine positions in the repeats. The translation termination signal is indicated with a star. Brackets show the sites of hybridization of oligo(dT) priming of different clones.

structure of a typical CEA cDNA clone, the nucleotide sequence, and the deduced amino acid sequence are shown in Fig. 1. The sequence of 2560 nucleotides includes a 5′ untranslated region of 97 base pairs (bp), a signal sequence of 102 bp, an amino-terminal domain of 321 bp which contains the known CEA amino-terminal amino acid sequence (see below), three strikingly similar repeating units of 534 bp each, and a short carboxy-terminal domain of 81 bp, followed by a 124-nucleotide 3′ untranslated region, a 213-nucleotide fragment of the repetitive Alu family, and a short poly(A) tail of 18 nucleotides.

Analysis by restriction mapping of the other cDNA clones revealed that most of them had the same 5′ end; this was probably because the nucleotide sequence of one of the probes used (a 54-mer) (Paxton et al., Int. Soc. Oncodev. Biol. Med. XIII 1985) was found in the 5′ end of the complete nucleotide sequence close to the amino-terminal codon of the mature CEA protein. Primer extension analysis of RNA preparations from the CEA-producing cell line confirmed that the cDNA clones were full length at the 5′ end; hybridization of a 21-mer oligonucleotide complementary to the 3′ end of the signal sequence and extension of this primer by reverse transcriptase with total RNA as a template produced a single band detected by denaturing gel electrophoresis which was 5 to 12 nucleotides longer than the 5′ end of three sequenced cDNA clones (Fig. 2).

Most of the clones also had the same 3′ end, terminating at the end of the Alu-like sequence, although four lacked most of the latter, giving a 3′ end 250 bp shorter, and two had a 3′ end which was about 500 bp longer. The most striking feature of 19 cDNA clones with the same 5′ and 3′ ends, however, was the central portion of each, which consisted of three, and occasionally four, very similar repeats. Eight additional cDNA clones which hybridized readily with all of the CEA-specific probes had different structures which could not be related to this basic repeat structure (data not shown). The nucleotide sequence of a three-repeat clone (Fig. 1) allows the conclusion that the repeats, while very similar, are not identical. This is consistent with the considerable evidence that the protein structure of CEA contains a series of repeated domains. In all cDNA clones examined, the repeats were in the same order.

Detailed structure of three-repeat CEA cDNA clones. Three separate three-repeat CEA cDNA clones were sequenced and found to have essentially identical sequences except for their 3′ ends, where one of the clones had a deletion of most of the Alu element. The 3′ end differences will be considered below. A composite nucleotide sequence and derived amino acid sequence of a three-repeat cDNA clone is given in Fig. 1B.

The 5′ untranslated region is characterized by the absence of any in-frame start or stop codons up to the first ATG in-frame codon found at position +1. This codon is in a context which conforms to the consensus of sequence of Kozak (20) for true translational initiation codons, being preceded 3 bp upstream by the canonical A and followed immediately downstream by a G. The 34-amino-acid signal sequence which follows demonstrates the usual features of a signal consensus sequence: a charged residue in the first few amino acids (Glu), a hydrophobic core region of nine amino acids, a helix breaker (Pro residues) and a small uncharged amino acid at the site of cleavage (Ala) (47).

The amino acid sequence of the amino-terminal end of CEA and its most prevalent normal counterpart, NCA, are known (10, 29, 35). Both sequences begin with Lys and are identical until position 55 (21 in mature protein), at which

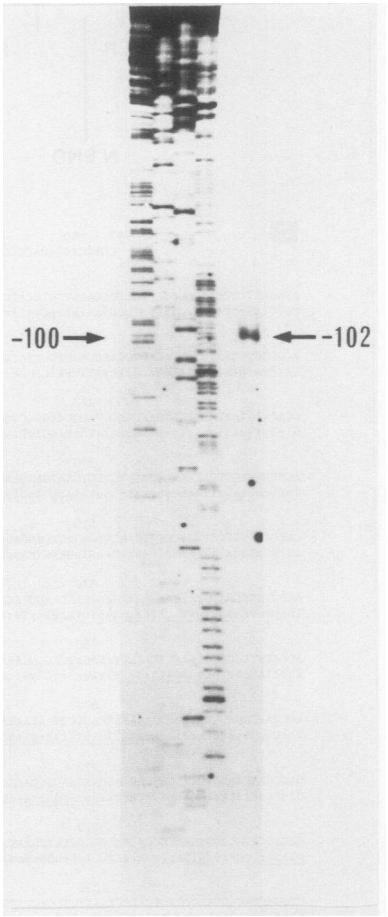

FIG. 2. Mapping the 5′ end of CEA mRNA from cell line LS180. A labeled 21-mer oligonucleotide complementary to the signal sequence was hybridized, as described in Material and Methods, to total RNA extracted from cell line LS180. Extension of this hybridized primer was done with avian myeloblastosis virus reverse transcriptase. The labeled extension product (right lane) was electrophoresed on a polyacrylamide sequencing gel, and the gel was subjected to autoradiography. A sequencing reaction was also performed on the 5′ end of a CEA cDNA M13 subclone by using the same 5′ ^{32}P-labeled primer. Numbers on the side correspond to positions of the DNA sequence of Fig. 1B.

CEA has Val and NCA has Ala. The predicted translated product of the nucleotide sequence starts with Lys immediately after the predicted cleavage point of the signal sequence and follows the known amino-terminal sequence exactly with Val at position 55 and with additional CEA-specific amino acids at positions 61, 62, 63, 77, 78, 80, 100, 123, and 139. This unequivocally identifies this cDNA as a CEA clone and not a NCA clone, since the cell line used to generate the cDNA library produces large amounts of CEA and not NCA, as demonstrated with CEA-specific immunoreagents (data not shown). The predicted amino acid sequence gives a total of 107 residues in the amino-terminal domain of the mature protein.

The core of the molecule is characterized by three similar repeat units of 534 bp coding for peptide domains of 178 amino acids. Alignments of these repeat units at the nucleotide level and at the protein level are presented in Fig. 3A and B, respectively. Analysis with the BESTFIT computer program revealed a degree of homology of 80.5 to 82% at the nucleotide level and 68 to 72.5% at the protein level,

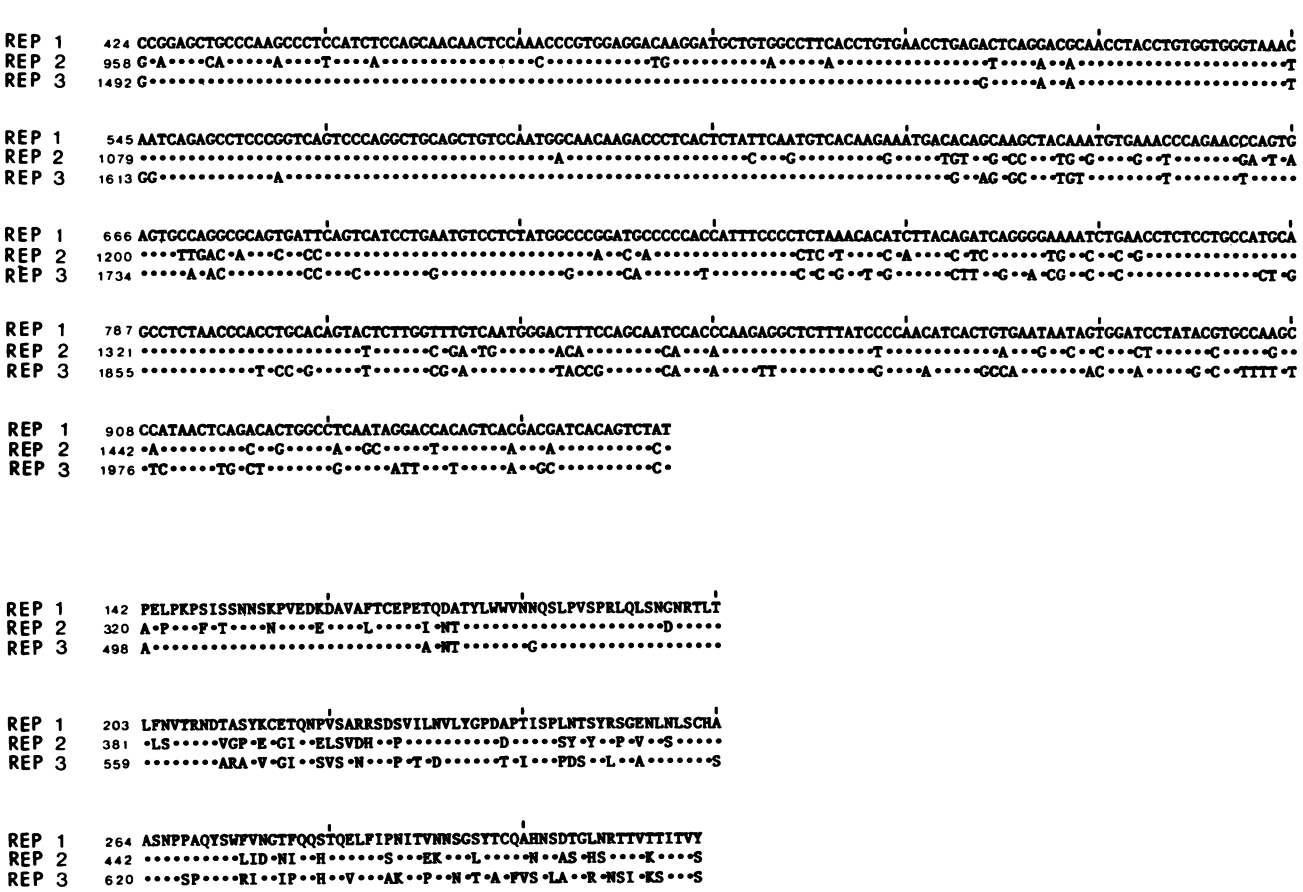

FIG. 3. Nucleotide and amino acid alignments of the three CEA repeats. (A) CEA repeats were aligned from nucleotide positions 424, 958, and 1492. Dots represent identical nucleotides; a printed nucleotide represents a substitution. (B) CEA repeats were aligned from amino acid positions 142, 320, and 498. Dots show identical residues; a printed amino acid indicates a substitution.

depending on the repeats compared. The higher degree of sequence alignment at the nucleotide level is highly unusual and will be considered in the Discussion. The protein alignment (Fig. 3B) also demonstrates the presence of an especially well conserved sequence defined by the beginning of the repeat units up to the second cysteine residue. Very few nucleotide changes occur in this region (22 of 200); those that lead to an amino acid replacement introduce only conservative changes respecting the size and charge of the original amino acid.

A 3′ carboxy-terminal domain characterized by a hydrophobic stretch of 18 nonpolar amino acids and 11 small slightly polar residues (Gly, Ser, and Thr) follows the repeats. This 27-amino-acid domain represents a possible membrane-anchoring region. The presence of a Met 11 amino acids from the end constitutes an interesting observation, since previous analyses of the amino acid composition of CEA indicated the complete absence of Met residues (34). A 3′ untranslated region of 124 nucleotides follows the translation stop codon (TAG) at position 2107. Of the 19 clones, 15 are terminated by a 213-nucleotide stretch representing a truncated human repetitive Alu sequence lacking 82 nucleotides of the usual 5′ end. The degree of homology between this Alu sequence and the consensus human Alu sequence is 90%. A second termination site was revealed by the isolation of four clones with similar 3′ end deletions. One of these, clone λ31, was sequenced and found to terminate 55 nucleotides beyond the 5′ end of the truncated Alu element, which was followed by a short poly(A) tail. Comparison of our sequence with that of Oikawa et al. (28), which includes an extra 479 nucleotides at the 3′ end, suggests that our clones were possibly initiated during cDNA synthesis by binding of the oligo(dT) primer to the A-rich region at the end of the Alu element for the longer clones, and to the A-rich region in the middle of the Alu element for the shorter ones. Second-strand synthesis of such cDNA molecules would result in perfect poly(A^+) tails of about 18 residues, as observed. One of our clones (clone λ17) was found to include the 479-bp 3′ end (data not shown). The absence of a typical poly$(A)^+$ addition signal in the nucleotide sequence of either the shortest or medium-length 3′ ends and the presence of such a signal in the longest 3′ end (28) favor the view that only the latter is truly represented in cellular mRNA. The preponderance of clones with medium-length 3′ ends (12 of 19), on the other hand, argues for the coexistence of a cellular mRNA corresponding to this termination site.

Homology of CEA to immunoglobulin supergene family. A computer-aided search of the National Biomedical Research Foundation protein databank (6) with the FASTP search program (43) using the entire CEA protein sequence indicated greatest homology to several members of the immunoglobulin supergene family as first reported by Paxton et al. (29) for partial CEA amino acid sequences. Our search discovered some new homologies, however, including rat immunoglobulin heavy chain (V region) (2), rat immunoglob-

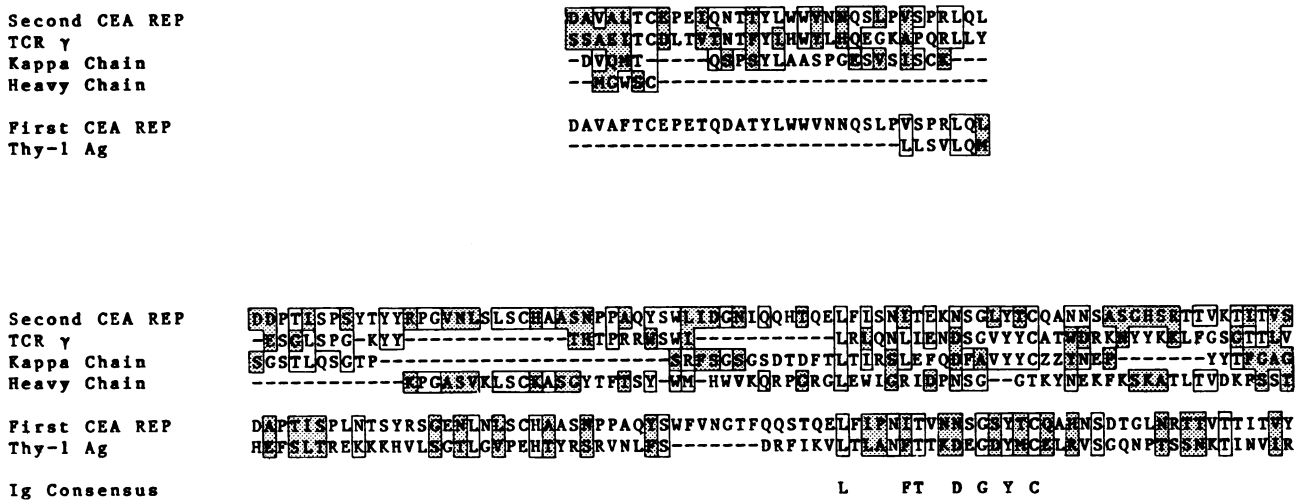

FIG. 4. Comparison of CEA repeat domains with immunoglobulin supergene family members. Amino acid comparison of the CEA repeats around cysteine residue 1 (top panel) and from before cysteine residue 3 to the end of the repeat (bottom panel) with four members of the immunoglobulin supergene family: human T-cell receptor γ chain (V region) (TCR γ); rat immunoglobulin light kappa chain (V region) (Kappa Chain); rat immunoglobulin heavy chain (V region) (Heavy Chain); rat Thy-1 antigen (Thy-1 Ag); and immunoglobulin supergene family consensus sequence of Williams and Gagnon (44) (Ig Consensus). Open boxes represent identical residues, stippled boxes show substitution with a conservative residue, and hyphens are gaps introduced to obtain maximum alignment.

ulin light kappa chain (V region) (39), rabbit poly-immunoglobulin receptor (27), mouse T-cell receptor β chain (C region) (12), and human T-cell receptor γ chain (V region) (9). These homologies were all clustered in the repeats of the CEA protein, with the strongest homology found in two distinct segments: the beginning of the CEA repeat up to the second cysteine and the region beside the cysteines 3 and 4. The first segment shows the least divergence between the repeats in the CEA protein itself (see above). The second segment corresponds to characteristic domains of all immunoglobulin chains in the vicinity of conserved cysteine disulfide loops. Of the three CEA repeats, the second shows the greatest degree of homology with the immunoglobulin supergene family.

To assess the significance of the homologies, we generated alignments between each of the CEA repeats and these proteins by using the ALIGN program of Dayhoff et al. (5). Gaps were introduced in the sequences to favor maximum alignments. Of all the proteins tested, only the human T-cell receptor γ chain presented an alignment score greater than 3.0 standard deviations away from random expectations, the minimum value considered to represent authentic relationships (8). This protein showed an alignment score of 8.1 standard deviations and 15% identical plus 22% conserved residues for a total of 37%. The actual alignments of the second CEA repeat to the human T-cell receptor γ chain, to the rat immunoglobulin kappa chain, and to the mouse immunoglobulin heavy chain and of the first CEA repeat to the primordial immunoglobulin supergene family member, rat Thy-1 antigen (26, 44), are shown in Fig. 4.

Identification of functional CEA cDNA clones. The complexity of the family of CEA cDNA clones emerging from the cDNA library of the LS180 cell line raises the question of whether all or just a subset of the clones are potentially functional. Poly(A)$^+$ RNA from LS180 and from various control lines was subjected to Northern analysis by using the insert of the three-repeat cDNA clone λ31 (which lacks most of the 3′ Alu sequence) as a radioactive hybridization probe. Under conditions of high hybridization stringency, the LS180 RNA showed three prominent bands, whereas a derived low-CEA-producing line (86/8) and human embryo fibroblasts showed none (Fig. 5), even after prolonged exposure (data not shown). These consisted of a major central band between two minor bands, each spaced by about 400 bp. Similar bands were seen with fresh clinical colon carcinoma samples (Fig. 5), indicating that they are not merely an artifact of this particular line. The identity of these bands is currently under investigation.

This result indicates the existence of mRNA species in cells corresponding to the cloned species of CEA cDNA, but does not prove that they all translated into cell surface

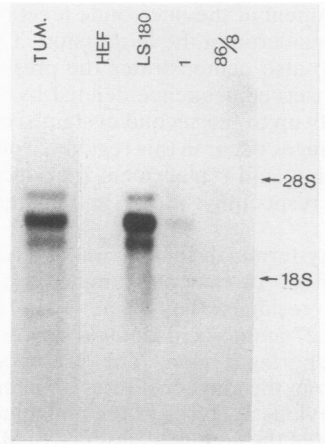

FIG. 5. Northern analysis of CEA mRNA from cell lines and from a human tumor. RNA was electrophoresed on a formaldehyde-agarose gel, transferred to a nitrocellulose membrane, and hybridized with the ^{32}P-labeled insert of clone λ31. After hybridization, the filter was washed as described in Materials and Methods and autoradiographed. Lanes: TUM., 10 μg of total RNA from colonic adenocarcinoma tissue; HEF, 1 μg of poly(A)$^+$ RNA from human embryo fibroblasts; LS180, 1 μg of poly(A)$^+$ RNA from cell line LS180; 1, 1 μg of poly(A)$^+$ RNA from medium CEA producer cell clone 1; 86/8, 1 μg of poly(A)$^+$ RNA from lowest-producer cell clone 86/8. 18S and 28S represent the positions of 18S and 28S rRNA.

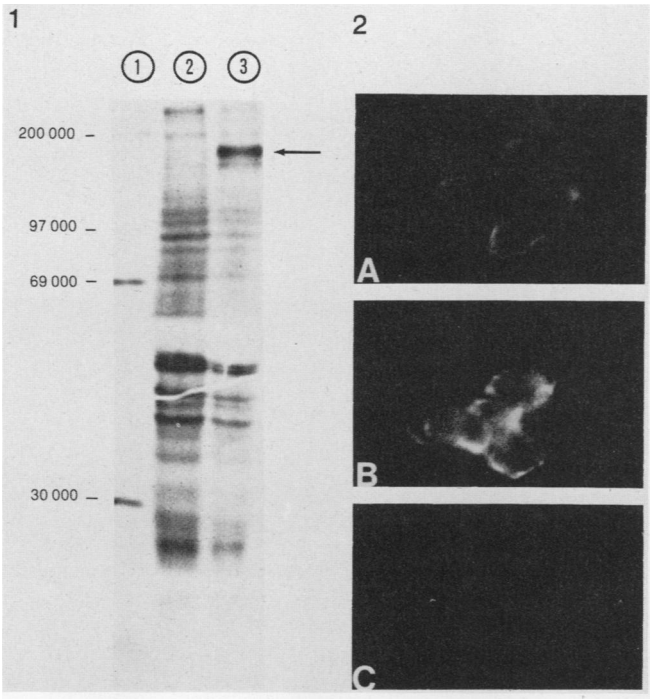

FIG. 6. Immunoprecipitation and immunofluorescence labeling of CEA on stable transfectants obtained with functional CEA cDNA. Panel 1, LR-73 cultures stably transfected (Table 1) with the functional CEA plasmid p91023B-31 were labeled with [^{3}H]leucine, solubilized, and immunoprecipitated with rabbit anti-CEA antibody (lane 3) or normal rabbit serum (lane 2). These were analyzed by electrophoresis on a sodium dodecyl sulfate-polyacrylamide gel along with labeled marker proteins (lane 1). The position of the immunoprecipitated CEA band in lane 3 is indicated by an arrow. From companion gels, this is the same position observed for glycosylated purified human CEA. (2) LR-73 cells, stably transfected with the functional plasmid p91023B-31 (panels A and B) or with the control antisense plasmid p91023B-27 (panel C), were suspended and incubated with rabbit anti-CEA antibody for 1 h at 4°C, washed with phosphate-buffered saline, and exposed to fluorescein-conjugated goat anti-rabbit immunoglobulin G for 30 min at 4°C. The antibody-treated cells were placed on microscope slides, and fluorescence micrographs were obtained by using Kodak Tri-X film (Eastman Kodak Co., Rochester, N.Y.).

proteins recognizable as CEA. Purified CEA has been reported to have a molecular weight of 180,000 with a protein component containing 12 cysteine residues and with a molecular weight of about 75,000 (34). This corresponds exactly to the predicted translation product of the three-repeat structure (12 cysteines, 72,813 molecular weight). In fact in vitro transcripts (30) of the three three-repeat structures with different 3′ ends subcloned in the bacterial expression vector pSP65 (25) were translated by using a cell-free rabbit reticulocyte system, and each yielded a single band of approximately 73,000 molecular weight, which was precipitable with specific anti-CEA polyclonal antibody (data not shown). The three three-repeat structures were also inserted into the mammalian expression vector p91023B and introduced into the CHO LR-73 cell line (Fig. 6; Table 1) and the monkey COS-1 cell line (data not shown) by the calcium phosphate procedure (14). After both transient and long-term stable assays, cell surface CEA was shown to be produced by the LR-73 transfectants by using the quantitative double monoclonal assay mentioned above (Table 1), by immunofluorescent staining of cells with anti-CEA monoclonal or polyclonal antibodies (Fig. 6, panel 2), and by immunoprecipitation of cellular proteins with anti-CEA polyclonal antibodies (Fig. 6, panel 1). The latter showed a band at the expected position for fully glycosylated human CEA (Fig. 6, panel 1). The level of CEA produced in the transfectants was as high as 175 ng/mg, or about 20% of the level seen in the high-producer cell line (Table 1). The variable levels of production seen with the different cDNA clones will be considered in the Discussion. All of the above tests were negative when applied to transfectants obtained with a CEA cDNA clone inserted into the vector in the antisense orientation (Fig. 6; Table 1).

DISCUSSION

We have isolated and characterized 27 CEA cDNA clones from the poly(A)$^+$ RNA of a human colon carcinoma cell line which produces large amounts of CEA (0.2% of the total cell protein). These clones were shown to contain the coding information for CEA by several criteria: their predicted translation product has an amino acid sequence which coincides exactly with that of CEA for 107 amino terminal residues and for 228 internal residues (compare Fig. 1 and Fig. 4 of reference 29 with Fig. 1 and Fig. 3 of this paper, respectively); their transcripts generated in vitro from SP-65 expression constructs can be translated in vitro into a polypeptide of the expected molecular weight which is precipitable by specific anti-CEA antibodies; and monkey and rodent cells transfected with expression constructs containing the cDNAs produce cell surface CEA identified by specific monoclonal antibodies. CEA is known to be one of a relatively large family of similar proteins (34). The evidence that our cDNA clones correspond to the CEA member of this family arises again from consideration of the amino acid sequence of their predicted translation product: not only does this agree exactly with that of CEA, but there are also 10 amino-terminal positions and 3 internal positions at which the sequence is different from that of the most prevalent normal counterpart of CEA, NCA-95 (29). Thus our cDNA clones do not correspond to NCA-95 and, in view of the cell type and organ specificity of other members of the CEA family (17), it seems very unlikely that they correspond to any other member of the family.

Of the 27 characterized CEA cDNA clones isolated from the library, 8 were found to be related by cross-hybridization

TABLE 1. Biological activity of CEA cDNA clones in animal cell expression vector[a]

DNA	CEA (ng/mg of extract protein)	
	Transient transfectants	Stable transfectants
Control	0.9	2.0
Antisense	0.1	2.5
p91023B-31	0.3	113
p91023B-25	0.2	22
p91023B-17	14	175

[a] LR-73 cells were transfected with 5 μg of the indicated DNA and 10 μg of CHO cell carrier DNA per culture by the calcium phosphate procedure (14). Stable transfectants were obtained by cotransfection with 0.5 μg of pSV2-AS per culture and selection with 2 mM albizziin (3). Assessment of antigen production was carried out at 48 h for the transient transfectants and at 2 to 3 weeks for pooled stable transfectants by using the double monoclonal antibody test (see Materials and Methods). The values shown for the transient transfectants represent the average of three experiments. Control DNA consisted of carrier DNA only; antisense and p91023B-31, p91023B-25, and p91023B-17 were CEA cDNAs inserted in antisense and sense orientations, respectively.

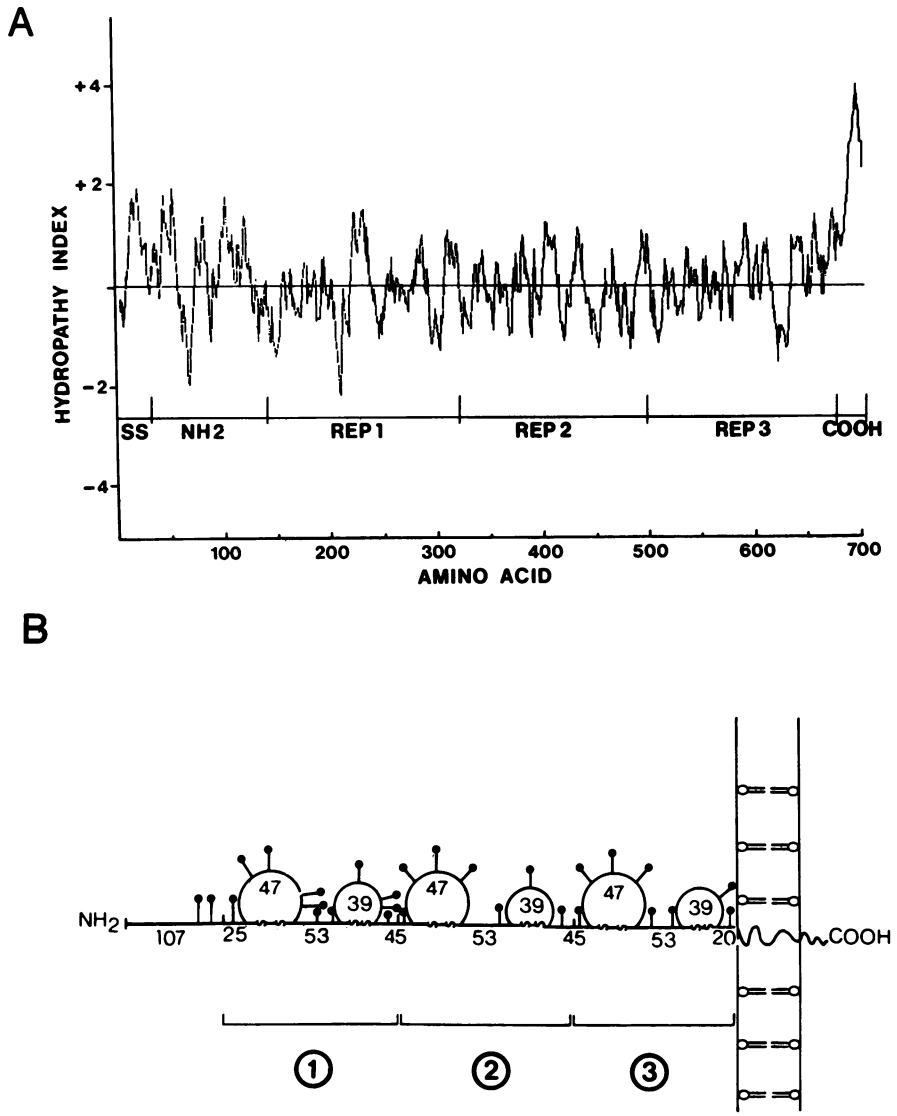

FIG. 7. Hydropathy plot (A) and schematic representation (B) of the putative structure of the CEA protein. The hydropathy plot was obtained by computer-assisted analysis with the algorithm and hydropathy values of Kyte and Doolittle (21) for a window of 13. Hydrophobic regions are above the line, and hydrophylic regions are below it. Abbreviations: SS, signal sequence; NH2, N terminus of mature protein; REP, repeat; COOH, C terminus of mature protein. The model in panel B shows predicted N-linked glycosylation sites (♆), transmembrane domain, and putative cysteine disulfide bridges. Circled numbers stand for repeats 1, 2, and 3. Other numbers denote the number of amino acids in each domain delineated by the amino terminus and the beginning of repeat 1, between the start of repeat 1 and cysteine residue 1 between each of the next 11 cysteines in the repeats, and finally between the last cysteine residue and the carboxy terminus of the protein.

with most of the other clones but to be quite different in their restriction maps; these will be the topic of a separate report. The rest of the clones had the same basic structure of a number (usually three) of repeats of 534 bp each embraced by common 5′ and 3′ ends. The 5′ ends were usually of the same length and started within a few bases of the 5′ start of transcription of the cellular CEA mRNA, whereas the 3′ ends fell into three classes: a few terminated at an A-rich region of a 5′-truncated repetitive Alu sequence, most terminated at the A-rich region at the end of the Alu element, and a few extended past the Alu element for a further 479 bp. The first two termination sites could represent an artifact of the technique of cDNA preparation, as mentioned in the Results. All three types of cDNA clones in an expression vector proved functional in animal cells, but the longest had the greatest activity, especially in transient assays (Table 1). This variable activity appears to be due, in these experiments at least, to variable transfection efficiencies.

Regarding the number of repeats in the CEA cDNA clones, it is clear that the majority (16 of 19) had three repeats and, from the nature and size of the predicted and demonstrated translation product, that this corresponds to the major form of CEA. A few of the clones, however, were observed to possess four repeats, a feature which appears not to be due to artifactual recombination between repeats during propagation. The observation by Paxton et al. (29) of a fourth amino acid sequence in one region of the repeat sequence supports the notion of a fourth repeat. Nucleotide sequencing of our four-repeat clone is in progress; until this information is in hand, it seems premature to speculate

further on the significance of variation in the number of repeats.

Analysis of the CEA amino acid sequence derived from the cDNA clones showed a protein with a processed hydrophobic signal sequence, the repeats with four cysteine residues each, all richly endowed with potential asparagine-linked glycosylation sites (28 in all), and a hydrophobic carboxy-terminal domain. A model for this structure, along with a hydropathy plot, is presented in Fig. 7, in which the carboxy-terminal domain is shown as a membrane anchor for the protein, consistent with the demonstrated localization of CEA to the external cell membrane (13, 16). From the demonstrated similarity between CEA and the rat Thy-1 antigen and various members of the immunoglobulin supergene family, we propose that cysteines 1 and 2 and cysteines 3 and 4 form two disulfide bridges, by analogy with bridges seen between analogous cysteines in the above proteins.

Does the present information shed any light on the function of CEA? Models for CEA function range between a stage-specific end product of development, in which case production in tumors would be adventitious, to a cell surface molecule involved in cellular growth control, in which production in tumors would be selected for. Two aspects of our results relate to function. The first concerns evolution. As with many other proteins, CEA has repeated domains, a strategy postulated to provide a means of developing further function in new domains while retaining the function of the original domain. For CEA, the repeat domains showed a greater similarity at the nucleotide sequence level (82%) than at the amino acid sequence level (69%), except for the first 67 amino acids, which showed little divergence. We speculate that these 67 amino acids specify a dominant function requiring integrity in every repeat, while the rest of the repeat domains determine a function requiring one valid copy only. The last appear to be in a state of early evolution, since the nucleotide sequences have diverged very little and the extent of the divergence of the amino acid sequences is precisely what would be predicted if nucleotide substitutions were random without selection for their effects on protein structure (as shown by probabilistic analysis [C. P. Stanners, unpublished data]).

The second functional implication arises from the comparisons with proteins of known function. The conserved portions of the repeats show significant homology with the V region of the human T-cell receptor γ chain, whereas the more variable portions show homology with other members of the immunoglobulin supergene family. It seems reasonable to propose, then, that the former conserved regions are involved in some dominant essential cell-cell recognition process, while the more diverged regions have some automatic consequential function. Disruption of this process by the production of large amounts of CEA (as much as 0.2% of the cell proteins, or 20% of the cell membrane proteins) in inappropriate situations could predispose the cell to uncontrolled growth, possibly initiated by oncogene activation. These speculations are being tested by molecular genetic studies of CEA evolution involving CEA-like proteins in other species and by studies on the function of our CEA-animal cell expression constructs introduced into various biological systems.

ACKNOWLEDGMENTS

We thank Aurora Labitan for technical assistance and Josie D'Amico for secretarial assistance. We are indebted to Randy Kaufman, Genetics Institute, for the use of the p91023B vector. We also thank Philippe Gros, Gordon Shore, and J. Pelletier for technical advice and Wendy Hauck and Lenore Beitel for their assistance in computer analysis.

N.B. is a recipient of a fellowship from the Fonds de Recherche en Santé du Québec. D.C. is supported by a fellowship from the National Cancer Institute of Canada. This work was supported by grants to A.F. and C.P.S. from the Medical Research Council of Canada and the National Cancer Institute of Canada.

LITERATURE CITED

1. **Aviv, J., and P. Leder.** 1972. Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid-cellulose. Proc. Natl. Acad. Sci. USA **69**:1408–1412.
2. **Bothwell, A. L. M., M. Paskind, M. Reth, T. Imanishi-Kari, K. Rajewsky, and D. Baltimore.** 1981. Heavy chain variable region contribution to the NP^b family antibodies: somatic mutation evident in a γ2a variable region. Cell **24**:625–637.
3. **Cartier, M., M. W. M. Chang, and C. P. Stanners.** 1987. Use of the *Escherichia coli* gene for asparagine synthetase as a selective marker in a shuttle vector capable of dominant transfection and amplification in animal cells. Mol. Cell. Biol. **7**:1623–1628.
4. **Chirgwin, J. M., A. E. Przybyla, R. J. MacDonald, and W. J. Rutter.** 1979. Isolation of biologically active ribonucleic acid from sources enriched in ribonucleases. Biochemistry **18**:5294–5299.
5. **Dayhoff, M. O., W. C. Barker, and T. L. Hunt.** 1983. Establishing homologies in protein sequences. Methods Enzymol. **91**: 534–545.
6. **Dayhoff, M. O., L. T. Hunt, W. C. Barker, B. C. Orcutt, L. S. Yeh, H. R. Chen, D. G. George, M. C. Blomquist, J. A. Fredrickson, and G. C. Johnson.** 1981. Protein sequence database. National Biomedical Research Foundation, Washington, D.C.
7. **Deverreux, J., P. Haeberli, and O. Smithies.** 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**:387–395.
8. **Doolittle, R. F.** 1981. Similar amino acid sequences: chance or common ancestry. Science **214**:149–159.
9. **Dyalynas, D. P., C. Murre, T. Quertermous, J. M. Boss, J. M. Leiden, J. G. Seidman, and J. L. Strominger.** 1986. Cloning and sequence analysis of complementary DNA encoding an aberrantly rearranged human T-cell γ chain. Proc. Natl. Acad. Sci. USA **83**:2619–2623.
10. **Engvall, E., J. E. Shively, and M. W. Wrann.** 1978. Isolation and characterization of normal crossreacting antigen (NCA). Homology of its NH_2-terminal amino acid sequence with that of carcinoembryonic antigen (CEA). Proc. Natl. Acad. Sci. USA **75**:1670–1674.
11. **Feinberg, A. P., and B. Vogelstein.** 1983. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. **132**:6–13.
12. **Gascoigne, N. R. J., Y. Chien, D. M. Becker, J. Kavaler, and M. M. Davis.** 1984. Genomic organization and sequence of T cell receptor β-chain constant- and joining-region genes. Nature (London) **310**:387–391.
13. **Gold, P., J. Krupey, and H. Ansari.** 1970. Position of the carcinoembryonic antigen of the human digestive system in ultrastructure of tumour cell surface. J. Natl. Cancer Inst. **45**: 219–222.
14. **Graham, F. L., S. Bachetti, R. McKinnon, C. P. Stanners, B. Cordell, and H. H. Goodman.** 1980. Transformation of mammalian cells with DNA using the calcium technique, p. 3–25. *In* R. Baserga, C. Croce, and G. Rovera (ed.), Wistar Symposium series, vol. 1, Introduction of macromolecules into viable mammalian cells. Alan R. Liss, Inc., New York.
15. **Gubler, U., and B. J. Hoffman.** 1983. A simple and very efficient method for generating cDNA libraries. Gene **25**:263–269.
16. **Haggarty, A., C. Legler, M. J. Krantz, and A. Fuks.** 1986. Epitopes of carcinoembryonic antigen defined by monoclonal antibodies prepared from mice immunized with purified carcinoembryonic antigen of HCT-8R cells. Cancer Res. **46**: 300–309.
17. **Hammarstrom, S., T. Svenberg, A. Hedin, and G. Sunblad.**

1978. Antigens related to carcinoembryonic antigen. Scand. J. Immunol. Suppl. 6 7:33–46.
18. **Hoyt, M. A., D. M. Knight, A. Das, H. I. Miller, and H. Echols.** 1982. Control of phage λ development by stability and synthesis of cII protein: role of the viral cIII and host hfl A, him A and him D genes. Cell **31**:565–573.
19. **Kaufman, R. J., and P. A. Sharp.** 1982. Construction of a modular dihydrofolate reductase cDNA gene: analysis of signals utilized for efficient expression. Mol. Cell. Biol. **2**:1304–1319.
20. **Kozak, M.** 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucleic Acids Res. **12**:857–872.
21. **Kyte, J., and R. F. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.
22. **Laemmli, U. K.** 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature (London) **227**:680–685.
23. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
24. **Maxam, A. M., and W. Gilbert.** 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA **74**:560–564.
25. **Melton, P. A., P. A. Krieg, M. R. Rebagliati, T. Maniatis, K. Zinn, and M. R. Green.** 1984. Efficient *in vitro* synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. Nucleic Acids Res. **12**:7035–7056.
26. **Moriuchi, T., H. C. Chang, R. Denome, and J. Silver.** 1983. Thy-1 cDNA sequence suggests a novel regulatory mechanism. Nature (London) **301**:80–82.
27. **Mostov, K. E., M. Friedlander, and G. Blobel.** 1984. The receptor for transepithelial transport of IgA and IgM contains multiple immunoglobulin-like domains. Nature (London) **308**: 37–43.
28. **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.
29. **Paxton, R. J., G. Mooser, H. Pande, T. D. Lee, and J. E. Shively.** 1987. Sequence analysis of carcinoembryonic antigen: identification of glycosylation sites and homology with immunoglobulin supergene family. Proc. Natl. Acad. Sci. USA **84**:920–924.
30. **Pelletier, J., and N. Sonenberg.** 1985. Insertion mutagenesis to increase secondary structure within the 5′ noncoding region of a eukaryotic mRNA reduces translational efficiency. Cell **40**: 515–526.
31. **Pollard, J. W., and C. P. Stanners.** 1979. Characterization of cell lines showing growth control isolated from both the wild type and leucyl-tRNA synthetase mutant of Chinese hamster ovary cells. J. Cell. Physiol. **98**:571–586.
32. **Rutzky, L. P., B. H. Tom, and B. D. Kahan.** 1984. Biological and antigenic analysis of human colon cancer cell clones. Prog. Cancer Res. Ther. **29**:135–145.
33. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
34. **Shively, J. E., and J. D. Beatty.** 1985. CEA-related antigens: molecular biology and clinical significance. Crit. Rev. Oncol. Hematol. **2**:355–399.
35. **Shively, J. E., M. J. Kessler, and C. W. Todd.** 1978. Amino-terminal sequences of the major tryptic peptides obtained from carcinoembryonic antigen by digestion with trypsin in the presence of Triton-X-100. Cancer Res. **38**:2199–2208.
36. **Shore, G. C., F. Power, M. Bendayan, and P. Carignan.** 1981. Biogenesis of a 35-kilodalton protein associated with outer mitochondrial membrane in rat liver. J. Biol. Chem. **16**:8761–8766.
37. **Shuster, J., D. M. P. Thomson, A. Fuks, and P. Gold.** 1980. Immunologic approaches to diagnosis of malignancy. Prog. Exp. Tumor Res. **25**:89–139.
38. **Stanners, C. P., G. L. Eliceiri, and H. Green.** 1971. Two types of ribosomes in mouse-hamster hybrid cells. Nature (London) **230**: 52–54.
39. **Starace, V., and P. Querinjean.** 1975. The primary structure of a rat κ Bence-Jones protein: phylogenetic relationships of V- and C-region genes. J. Immunol. **115**:59–62.
40. **Terry, W., P. Henkart, J. Coligan, and C. Todd.** 1974. Carcinoembryonic antigen: characterization and clinical applications. Transplant. Rev. **20**:100–129.
41. **Thompson, J. A., H. Pande, R. J. Paxton, L. Shively, A. Padma, R. L. Simmer, C. W. 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.
42. **Wertman, K. F., A. R. Wyman, and D. Botstein.** 1986. Host/vector interactions which affect the viability of recombinant λ clones. Gene **49**:253–262.
43. **Wilbur, W. J., and D. J. Lipman.** 1983. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. USA **80**:726–730.
44. **Williams, A. F., and J. Gagnon.** 1982. Neuronal cell Thy-1 glycoprotein: homology with immunoglobulin. Science **216**: 696–703.
45. **Wong, G. G., J. S. Witek, P. A. Temple, K. M. Wilkins, A. C. Leary, D. P. Luxenberg, S. S. Jones, E. L. Brown, R. M. Kay, E. C. Orr, C. Shoemaker, D. W. Golde, R. J. Kaufman, R. M. Hewick, E. A. Wang, and S. C. Clark.** Human GM-CSF: molecular cloning of the complementary DNA and purification of the natural and recombinant proteins. Science **228**:810–814.
46. **Wood, W. I., J. Gitschier, L. A. Lasky, and R. M. Lawn.** 1985. Base composition-independent hybridization in tetramethylammonium chloride: a method for oligonucleotide screening of highly complex gene libraries. Proc. Natl. Acad. Sci. USA **82**: 1585–1588.
47. **Yamamoto, T., C. G. Davis, M. S. Brown, W. J. Schneider, M. L. Casey, J. L. Goldstein, and D. M. Russell.** 1984. The human LDL receptor: a cysteine-rich protein with multiple Alu sequences in its mRNA. Cell **39**:27–38.
48. **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.