

PRIMARY STRUCTURE OF HUMAN CARCINOEMBRYONIC ANTIGEN (CEA)  
DEDUCED FROM cDNA SEQUENCE

Shinzo Oikawa<sup>1</sup>, Hiroshi Nakazato<sup>1</sup>, and Goro Kosaki<sup>2</sup>

<sup>1</sup>Suntory Institute for Biomedical Research, Shimamoto-cho,  
Mishima-gun, Osaka 618, Japan

<sup>2</sup>Tokyo Metropolitan Komagome Hospital, Bunkyo-ku, Tokyo, 113, Japan

Received December 5, 1986

---

The cDNAs corresponding to the mRNA encoding a polypeptide which is immunoreactive with the antisera specific to carcinoembryonic antigen (CEA) (1) are cloned. The amino acid sequences deduced from the nucleotide sequences of the cDNAs show that it is synthesized as a precursor with a signal peptide followed by 668 amino acids of the putative mature CEA peptide, whose N-terminal 24 amino acids and amino acids 286 to 295 exactly coincide with those known for N-terminal sequences of CEA (2) and NFA-1 (3), respectively. The first 108 N-terminal residues are followed by three very homologous repetitive domains of 178 residues each and then by 26 mostly hydrophobic residues which probably comprise a membrane anchor. Each repetitive domain contains 4 cysteines at precisely the same positions and as many as 28 possible N-glycosylation sites are found in the CEA peptide region agreeing with high carbohydrate content of purified CEA. © 1987 Academic Press, Inc.

---

CEA first described by Gold and Freedman (1) is a complex immunoreactive glycoprotein with a molecular weight of 180,000 comprising 60% carbohydrate (4) found in adenocarcinomas of endodermally derived digestive system epithelia and foetal colon. The roles of the CEA immunoassay for diagnosis and serially monitoring cancer patients for recurrent disease or response to therapy, particularly in the case of colonic cancer, have been widely evaluated and documented. But in practicing the immunoassay for such purposes, one encounters difficulties because of the presence of CEA-like substances e.g. NCA or NCA-1 (5), NCA-2 (6), TEX (8), various normal faecal antigens (NFA-1, NFA-2 and NFCA) (9) and BGP-I (10), in normal colonic mucosa and in the serum, saliva, faeces and colon lavages of apparently normal individuals. These CEA-like substances share with CEA antibody binding sites which are mostly on peptide moieties in addition to distinct binding sites suggesting

---

**Abbreviations:** CEA, carcinoembryonic antigen; NCA, nonspecific crossreacting antigen; TEX, tumor-extracted CEA-related antigen; NFA, normal faecal antigen; NFCA, normal faecal cross-reacting antigen; BGP-I, biliary glycoprotein-I; b, base; kb, kilobase; bp, base pair.

the presence of very closely related genes as well as precursor-product relationships between some of them (11). It is essential to clarify the molecular nature of these antigens for establishing the tumor-specific assays for clinical use and for the study of their biological significance.

In this report we describe the cloning of cDNAs corresponding to the mRNA encoding a polypeptide which is immunoreactive with the antisera specific to CEA and the primary structure of the precursor for the putative CEA deduced from the nucleotide sequences of the cDNAs.

## METHODS

**cDNA Cloning.** Poly(A)<sup>+</sup> RNA was prepared from total RNA extracted from 1.3g of human malignant colon tissue from a patient as described (16), and a cDNA library was constructed using a phage expression vector  $\lambda$ gt11 (17). Double stranded DNA corresponding to the poly(A)<sup>+</sup> RNA was prepared according to the RNaseH method of Watson et al. (18), methylated at *Eco*RI sites, ligated to *Eco*RI linkers, digested by *Eco*RI, and size fractionated on a Bio-Gel A-50m column. The cDNA larger than 500-bp was ligated to the *Eco*RI-cut dephosphorylated  $\lambda$ gt11 DNA (Promega Biotec, U.S.A.) and packaged using Gigapack (Vector Cloning Systems, U.S.A.). Approximately  $3 \times 10^5$  clones were plated and submitted to immunological screening. One positive clone termed  $\lambda$ Kr40 was detected and purified by plaque purification. A cDNA insert was isolated by electrophoresis following cleavage with *Eco*RI and ligated to pBR322. Resulting plasmid, pBRCEA5 was used to transform *E. coli* DH1 to prepare DNA for further study. Two *Pvu*II fragments (from 566 to 1099 and from 1100 to 1633, Fig.2) were labelled by nicktranslation with [ $\alpha$ -<sup>32</sup>P]dCTP and used as a hybridization probe in RNA blot analysis shown in Fig.4 and in the screening of a cDNA library prepared from poly(A)<sup>+</sup> RNA isolated from healthy part of the colon tissue of the same patient as above. After sucrose density gradient centrifugation, poly(A)<sup>+</sup> RNA of 2 to 5-kb was pooled and used for the construction of a cDNA library by the method of Okayama and Berg (19). Approximately 7,000 clones of *E. coli* DH1 transformants thus obtained were transferred to nitrocellulose filters and screened. After hybridization with the probes described above for 16hr at 65°C, 8 positive clones, of which 4 contained longer cDNA inserts than  $\lambda$ Kr40 were obtained. After restriction enzyme mapping, plasmids, pCEA55-2 and pCEA8-11 were isolated for nucleotide sequence analysis as described in Fig.2 legend.

**Immunological Methods.** Immunological staining of plaques transferred to nitrocellulose membranes was performed according to the manufacturers instruction (Express-Blot Assay Kit, Bio-Rad Lab. U.S.A.). The first antibodies for screening and for the confirmation of the immunoreactivities were rabbit anti-CEA antibody (DAKOPATTS, Denmark) and those described in Fig.1 legend, respectively. Peroxidase labelled goat anti-rabbit IgG antibody and rabbit anti-sheep IgG antibody are from Bio-Rad Lab. (U.S.A.) and Kirkegaard Perry Lab. Inc. (U.S.A.), respectively.

**Blot Hybridizations.** RNA was separated by electrophoresis in a 1% agarose gel after denaturation with formamide and transferred to nitrocellulose sheet. Hybridization with the <sup>32</sup>P-labelled probe described above was done for 16hr at 45°C. Ten  $\mu$ g of DNA was digested with 50units of endonuclease for 3hr at 37°C and electrophoresed, transferred to nitrocellulose sheet (25), hybridized for 17hr at 65°C with N-terminal 201-bp *Pst*I-*Pvu*II fragment of pCEA 80-11 insert which had been labelled by nick translation with [ $\alpha$ -<sup>32</sup>P]dCTP.

## RESULTS AND DISCUSSIONS

A cDNA library was constructed using poly(A)<sup>+</sup> RNA extracted from colon tumor of a patient and *Eco*RI-cut dephosphorylated  $\lambda$ gt11 DNA and plated on *E.coli* as described in Fig.1 legend. A plaque was positively stained in the immunoscreening assay using a rabbit anti-CEA antibody as the first antibody and the cloned phage termed  $\lambda$ Kr40 was purified by plaque purification. To make sure the specificity of the immunoreaction we conducted the immunostaining of the plaques formed by plating  $\lambda$ Kr40 on *E.coli* with anti-CEA antibodies of different sources. As is shown in Fig.1, all four antibodies including NCA- or NCA-2- adsorbed antibody but not non-immune serum recognized the  $\lambda$ Kr40 encoded product indicating that  $\lambda$ Kr40 carries the cDNA which encodes CEA.

As the nucleotide sequence of the cDNA insert determined as shown in Fig.2 indicated the absence of N-terminal sequence corresponding to that known for CEA (2), Okayama-Berg cDNA library was prepared using poly(A)<sup>+</sup> RNA isolated from apparently normal colon tissue removed from the same patient as above. The tissue was chosen, for RNA blot hybridization analysis had shown abundant existence of RNAs which hybridized with the mixture of two 534-bp *Pvu*II fragments isolated from the  $\lambda$ Kr40 cDNA insert (lane 5 in Fig.4A). Using the same probe for a colony hybridization screening, two clones of plasmids, pCEA55-2 and pCEA80-11, which have cDNA inserts longer than that in  $\lambda$ Kr40 were isolated (Fig.2).

The nucleotide sequences determined by the analyses of the three clones (Fig.2) reveal a long open reading frame followed by 3'-non-coding region of at least two kinds. The cDNA in the clone pCEA80-11 has a poly(A) stretch 21-b downstream of AATAAA sequence whereas that of the clone pCEA55-2 has a poly(A) stretch about 500-b further down stream suggesting the existence of different mode of transcription or processing of the primary transcript. The

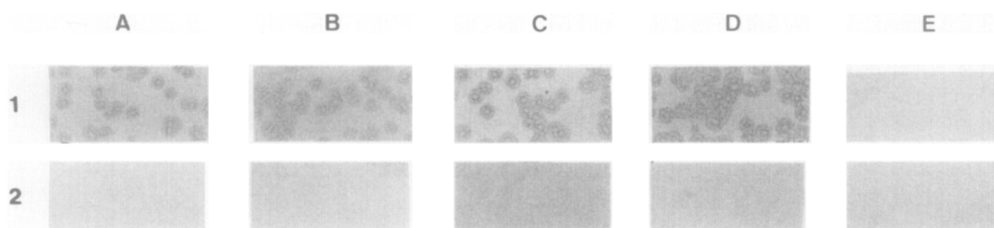


**Fig.1.** Reactivity of  $\lambda$ Kr40 product to antibodies of different sources. Plaques formed by 1;  $\lambda$ Kr40 or 2;  $\lambda$ gt11 plated on *E.coli* Y1090 were transferred and fixed to nitrocellulose membrane which is then cut into pieces, and reacted with anti-CEA antibodies of different animals and sources before staining as described in METHODS. A, rabbit (A115, DAKOPATTS, Denmark); B, sheep, NCA-2 adsorbed (640733-1, Miles, U.S.A.); C, sheep (AHP 105A, Serotec, England); D, sheep, NCA adsorbed (AHP 105B, Serotec); E, sheep, non-immune serum.

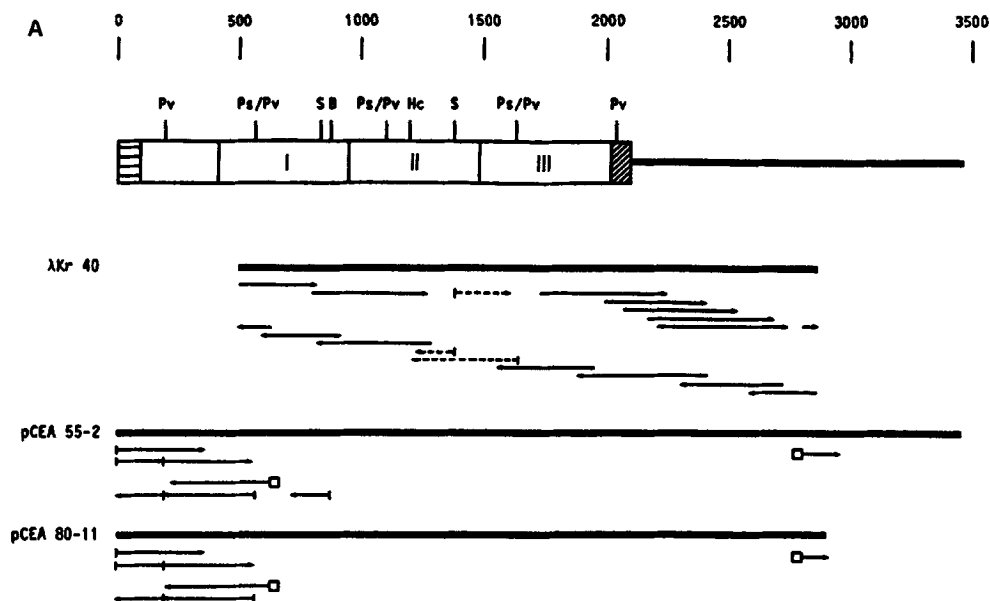


**Fig.2.** Sequencing strategy of cloned cDNA (A), and primary structures of cDNA and the CEA (B). In A, schematic structure of the cDNA depicting protein-coding (open boxes) and non-coding regions (■) is shown at the top. The restriction endonuclease sites only relevant to the present work are shown. Three cDNA inserts are shown by thick horizontal bars. The direction and extent of sequence determinations are shown by horizontal arrows under each clone used. ■, signal peptide; □, N-terminal 108-residue sequence; I, II, III, domain I, II, III; ▨, hydrophobic C-terminus. Restriction endonuclease and 5'-terminal nucleotide generated by cleavage are: Pv, PvuII (187, 566, 1100, 1634 & 2043); Ps, PstI (565, 1099 & 1633); S, SacI (841 & 1375); B, BamHI (873); Hc, HincII (1193). In B, nucleotide residues are numbered in the 5' to 3' direction, beginning with the first residue of the insert in pCEA80-11, the number of the nucleotide residue at the right end of each line is given. The deduced amino acid sequence is shown below the nucleotide sequence, and amino acid residues are numbered so that N-terminal sequence coincides with that reported for CEA. Residues of the signal peptide are indicated by negative numbers. ●, start of pCEA55-2 insert; o, start and end of λKr40 insert; \*, translational stop; †, start of poly(A) in pCEA80-11. Truncated Alu family sequence is underlined and the direct repeats boxed. N-terminal 24 amino acids identical to those of CEA are boxed; ★, the residue is alanine for NCA-1. Amino acids identical to those of N-terminus of NFA-1 are underlined. Starts and ends of domain I, II and III are shown above the nucleotide sequence. C-terminal hydrophobic region is shown by broken underline. Nucleotide sequencing of λKr40 insert was done by two methods. 1), a 2,400-bp EcoRI restriction fragment of the insert was recloned into M13mp18 and M13mp19 and sequenced (20) (→); 2), appropriate restriction fragments shown in the figure were recloned into M13mp18 or M13mp19, and sequenced by the chain termination method (21) (---→). pCEA55-2 and pCEA80-11 were analysed by 1), chain termination method on appropriate restriction fragments recloned into pUC9 (22) using general and reverse primers (→); 2), chain termination method using two synthetic oligodeoxyribonucleotide primers 5'-GTTTCACATTGTAGCT-3' and 5'-ATGTGGTCGCTCCAGAC-3' (□→).

analysis of this region is left for future study. Another feature to be noted about the 3'-non-coding region is that it contains an Alu family sequence (12) flanked by a pair of direct repeats. Compared to the Alu family consensus sequence (12), 71 nucleotides are deleted from 5'-terminal region of the left half of the dimer.

|          |                                                                                                                                                                                                                                                                             |      |
|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| <b>B</b> | 5'.....CTCGGCCCCCTCCACAGATGGTGCATCCCTGGCAGAGGCTCTGCTCAGACCTCACTTCTAACCTTCGGAACCCGCCACCACTGCC<br>SerAlaProHisArgTrpCys11eProTrpGlnArgLeuLeuLeuThrAlaSerLeuLeuThrPheTrpAsnProProThrThrAla<br>-30 -20 -10 -1                                                                   | 91   |
|          | <u>AAAGCTCACTATTGAATTCACGGCGTTCATGTCGACAGGGGAGGAGTCTTCTACTTGTCCCAATCTGCCCCAGCATCTTTTGGCTACAGCTGGTACAAAGGTGAAGAGTGGAT</u><br><u>LysLeuThr1eGluSerThrProPheAsnVal1AlaGluGlyLysGluValLeuLeuLeuValHisAsnLeu</u> ProGlnHisLeuPheGlyTyrSerTrpTyrLysGlyGluArgValAsp<br>10 20 30 40 | 211  |
|          | GGCAACCGTCAAATTATAGGATATGTAATGAAGTCAACAAGCTACCCAGGGCCGCGATGCTGCTGAGAGATAATATACCCCAATGCATCCCTGCTGATCCAGAACATCATCCAG<br>GlyAsnArgGln1e1eGlyTyrVal11eGlyThrGlnGlnAlaThrProGlyProAlaTyrSerGlyArgGlu11e1eTyrProAsnAlaSerLeuLeu11eGlnAsn11e1eGln<br>50 60 70 80                   | 331  |
|          | AATGACACAGGATTCTACACCTACAGCTATAAGTCAGATCTTGTGAATGAAGAAGCAACTGGCCAGTTCGGGATATACCGGAGCTGCCAAGCCCTCCATCTCCAGCAACAACCTCC<br>AsnAspThrGlyPheTyrThrLeuHisVal11eLysSerAspLeuValAsnGluGluAlaThrGlyGlnPheArgValTyrProGluLeuProLysProSer11eSerSerAsnAsnSer<br>90 100 110 120          | 451  |
|          | AAACCCGTGGAGGACAAGGATGCTGGGCTTCACCTGTGAACCTGAGACTCAGGACGCAACCTACCTGTGGTGGGTAACCAATCAGAGCCTCCGGTCAGTCCAGGCTGAGCTGTCC<br>LysProVal1GluAspLysAspAlaVal1AlaPheThrCysGluProGluThrGlnAspAlaThrTyrLeuTrpTrpValAsnAsnGlnSerLeuProVal1SerProArgLeuGlnLeuSer<br>130 140 150 160       | 571  |
|          | AATGGCAACAGGACCTCTACTTATCAATGTCACAAGAAATGACACAGCAAGCTACAAATGTBAACCCAGAACCCAGTGAGTGCAGGCGCAGTATTGATCATCTGATGTCTC<br>AsnGlyAsnArgThrLeuThrLeuPheAsnVal1ThrArgAsnAspThrAlaSerTyrLysCysGluThrGlnAsnProVal1SerAlaArgArgSerVal11eLeuAsnVal1Leu<br>170 180 190 200                 | 691  |
|          | TATGGCCGGATGCCCCACCATTTCCCTCTAAACACATCTTACAGATCAGGGGAAATCTGAACCTCTCTGCCACGACGCTCTAACCCACTGCACAGTACTCTTGGTTGTCAAT<br>TyrGlyProAspAlaProThr11eSerProLeuAsnThrSerTyrArgSerGlyGluAsnLeuAsnLeuSerCysHisAlaAlaSerAsnProProAlaGlnTyrSerTrpPheVal1Asn<br>210 220 230 240            | 811  |
|          | GGGACTTTCAGCAATCCACCAAGAGCTCTTATCCCAACATCACTGTGAATAATAGTGGATCCTATACGTGCAAGCCCACTACTCAGACACTGGCTCAATAGACCACAGTCAAG<br>GlyThrPheGlnGlnSerThrGlnGluLeuPhe11eProAsn11eThrValAsnAsnSerGlySerTyrThrCysGlnAlaHisAsnSerAspThrGlyLeuAsnArgThrThrVal1Thr<br>250 260 270 280           | 931  |
|          | ACGATCACAGTCTATGACAGGACCCCAACCCCTTCATCACCAGCAACAACCTCAACCCCGTGGAGGATGAGGATGCTGTAGCCTTAACCTGTGAACCTGAGATTGAGAACCAACCTAC<br>Thr11eThrVal1TyrAlaGluProProLysProPhe11eThrSerAsnAsnSerAsnProVal1GluAspGluAspAlaVal1AlaLeuThrCysGluProGlu11eGlnAsnThrThrTyr<br>290 300 310 320    | 1051 |
|          | CTGTGGTGGGTAATATCAGAGCCTCCGGTCAGTCCAGGCTGCAGCTGTCCAATGACAAAGGACCTCACTCTACTAGTGTCAAGGAATGATGTAGGACCTTATGAGTGTGGA<br>LeuTrpTrpVal1AsnAsnGlnSerLeuProVal1SerProArgLeuGlnLeuSerAsnAsnAspThrLeuThrLeuLeuSerVal1ThrArgAsnAspVal1GlyProTyrGluCysGly<br>330 340 350 360             | 1171 |
|          | ATCCAGACGAATTAGTGTGACCAAGCAGGACCTCATCTGAATGTCTCTATGGCCGACGACCCACCATTTCCCTCCATACACCTATTACCGTCCAGGGGTGAACCTCAGC<br>11eGlnAsnGluLeuSerValAspHisSerAspProVal11eLeuAsnVal1LeuTyrGlyProAspAspProThr11eSerProSerTyrThrTyrArgProGlyVal1AsnLeuSer<br>370 380 390 400                 | 1291 |
|          | CTCTCTGCCATGACGCTCTAACCCACCTGCACAGTATTCTTGGCTGATTGATGGGAACATCCAGCAACACACAGAAGCTCTTATCTCCAACTCACTGAGAAGACAGCGGACTC<br>LeuSerCysHisAlaAlaSerAsnProProAlaGlnTyrSerTrpLeu11eAspGlyAsn11eGlnGlnHisThrGlnGlnLeuPhe11eSerAsn11eThrGluLysAsnSerGlyLeu<br>410 420 430 440            | 1411 |
|          | TATACCTGCCAGGCCAATAACTCAGCCAGTGGCCAGCAGGACTACAGTCAAGACAATCAGACTCTCTGCGAGCTGCCAAGCCCTCCATCTCCAGCAACAACCTCAAACCCGTGGAG<br>TyrThrCysGlnAlaAsnAsnSerAlaSerGlyHisSerArgThrThrVal1LysThr11eThrVal1SerAlaGluLeuProLysProSer11eSerSerAsnAsnSerLysProVal1Glu<br>450 460 470 480      | 1531 |
|          | GACAAGGATGCTGTGGCTTCACTGTGAACCTGAGGCTCAGAACCAACCTACCTGTGGTGGGTAATGTCAGAGCCTCCAGTCACTCCAGGCTGCAGCTGTCCAATGGCAACAGG<br>AspLysAspAlaVal1AlaPheThrCysGluProGluAlaGlnAsnThrThrTyrLeuTrpVal1AsnGlyGlnSerLeuProVal1SerProArgLeuGlnLeuSerAsnGlyAsnArg<br>490 500 510 520            | 1651 |
|          | ACCCTCACTCTATTCAATGTCAAGAATGACGCAAGGCTATGATGTGAATCCAGAACTCAGTGTGCAACCCAGTGCACCTGATGTCTCTATGGGCGGAC<br>ThrLeuThrLeuPheAsnVal1ThrArgAsnAspAlaArgAlaTyrVal1CysGly11eGlnAsnSerVal1SerAlaAsnArgSerAspProVal1ThrLeuAspVal1LeuTyrGlyProAsp<br>530 540 550 560                      | 1771 |
|          | ACCCCATCATTTCCCCCAGACTGCTTACCTTTCCGGAGCGAACCTCAACCTCTCTGCCACTGGGCTCTAACCCATCCCGCAGTATTCTTGGGCTATCAATGGGATACCGCAG<br>ThrPro11e11eSerProProAspSerSerTyrLeuSerGlyAlaAsnLeuAsnLeuSerCysHisSerAlaSerAsnProSerProGlnTyrSerTrpArg11eAsnGly11eProGln<br>570 580 590 600             | 1891 |
|          | CAACACACAAAGTCTCTTTATCGCCAAATCAGCGCAATAATAACGGACCTATGCTGT11eTGTCTAACTTGGCTACTGGCGCAATAATCCATAGTCAAGGATCAGAGTC<br>GlnHisThrGlnVal1LeuPhe11eAlaLys11eThrProAsnAsnAsnGlyThrTyrAlaCysPheVal1SerAsnLeuAlaThrGlyArgAsnAsnSer11eVal1LysSer11eThrVal1<br>610 620 630 640            | 2011 |
|          | TCTGCATCTGGAATCTCTCTGGTCTCAGCTGGGCGCACTGTGGCATCATGATTGAGAGTGTGGTGGGTTGCTCTGATATAGCAGCCCTGGTGTAGTTTCTTATTTCAGGAAGA<br>SerAlaSerGlyThrSerProGlyLeuSerAlaGlyAlaThrVal1Gly11eMet11eGlyVal1LeuVal1GlyVal1AlaLeu11e<br>650 660 668                                                | 2131 |
|          | CTGACAGTTGTTTGTCTTCTCTTAAAGCATTGCAACAGCTACAGTCAAAATTGCTCTTTACCAAGGATATTTACAGAAAGACTCTGACGAGAGATCGAGACCATCTAGCCAAC                                                                                                                                                           | 2251 |
|          | ATCGTGAACCCCATCTCTACTAAAAATACAAAAATGAGCTGGGCTTGGTGGCGCCACTGTAGTCCAGTTACTCGGAGGCTGAGGAGGAGAAATCGCTGAACCCCGGAGGTGGAG                                                                                                                                                          | 2371 |
|          | ATTGCACTGAGCCAGATCGCACCCTGACCTCCAGCTTGCGCAACAGAGCAAGACTCATCTCAAAAAGAAAAGAAAAGAGACTCTGACCTGTACTCTTGAATACAAGTTTCTGATACC                                                                                                                                                       | 2491 |
|          | ACTGCACTGTCTGAGAATTTCCAAAATTTAATGAATTAAGTACAGCTTCATGAACTGTCCACCAAGATCAAGCAGAGAAAAATTAATTTATCGGACTAAATGAATTAAGGAGGA                                                                                                                                                          | 2611 |
|          | TAATATTTTATAATTTTATTTGAAATTTGCTGATCTTTAAATGTCTTGTTCGCCAGTTTCAGGAACTTTTTTTCTTTTAAAGCTATCCACAGCTTACAGCAATTTGATAAAAATA                                                                                                                                                         | 2731 |
|          | TACTTTTGTGAACAAAAATGAGACATTACATTTTCTCCCTATGTGGTGCCTCCAGCTTGGGAACTATTTCATGAATATTATATGTATGTGTAATATAGTTATGACAGTTCAAT                                                                                                                                                           | 2851 |
|          | AAAATCTGCTCTTGTATGACAGAAATACATTTGAAAAATTGGTTATATTACCAAGACTTTGACTAGAATGTCGTAT.....3'                                                                                                                                                                                         | 2929 |

Fig. 2—Continued.

The long open reading frame truncated at its N-terminal portion including the initiative methionine contains 2095 nucleotides encoding 698 amino acids. The first 30 amino acids seem to comprise a truncated signal peptide (13) which generally precedes secretory proteins and contains a hydrophobic core (14). Alanine which is often found at the end of many signal sequences (14) also marks the end of the signal sequence in this case and is followed by a protein of 668 amino acids, which most likely is the mature form of CEA peptide for the following reasons in addition to the immunoreactivities (Fig.1). First, the N-terminal 24 amino acids are exactly identical to those reported for CEA (2) and differ from those of NCA-1 (7) and TEX (8) at 21st residue which is valine for the former and alanine for the latters. Second, it contains only one methionine at position 658, *i.e.* methionine comprises 0.15% of total residues. The low content of methionine is in good agreement with that reported by Krantz and Laferte (15) and distinguish CEA from NCA-1 which are reported to have small but a definite quantity of methionine when little, if any, is found in CEA (2), NFA-1 and NFA-2 (3). Finally, amino acid sequence identical to N-terminal 14 amino acids of a fragment released from a CEA preparation (X556, DAKOPATTS, Denmark) treated with a mixture of trypsin (Sigma, U.S.A.) and *Achromobacter* protease I (Wako, Japan) is found at the positions 49 to 62 in the protein (Fig.2B, experimental details will be published elsewhere). Accordingly, we tentatively call the 668-residue peptide CEA.

The N-terminal 108 amino acids are followed by three homologous repeats of 178 residues named domain I, II and III, the amino acid sequence homology between I and II, II and III, III and I and I, II and III is 73.0, 67.4, 72.5, and 61.8%, respectively (Fig.2 & Fig.3). All 12 cysteine residues are found in this region, each domain having 4 residues at the same positions as others. The finding of a sequence identical to the N-terminal 10 amino acids of NFA-1 at positions 286 through 295 suggests that NFA-1 (3) is the degradation product of CEA. It is interesting to note in this context that the domain II somewhat diverges from the domain I and III in this region (Fig.3). The domain III is followed by 26 mostly hydrophobic amino acids which seem to form a membrane anchor region. There are 28 possible N-glycosylation sites, *i.e.* Asn-X-Thr/Ser (X is any residue except Pro), 2 in N-terminal sequence, 11, 7 and 8 in the domain I, II and III, respectively. The peptide is rather acidic and hydrophilic except for the C-terminal short sequence.

Figure 4A depicts the blot hybridization analysis of RNA from human cells and tissues using as a probe  $^{32}\text{P}$ -labelled *Pvu*II fragments which correspond to the repetitive domains of the CEA (Fig.2). The presence of multiple size species of RNA, *i.e.* major 3.5-kb species and minor 4.2-, 2.9- and 1.6-kb species, reflects partly the fact that two cDNA clones which have

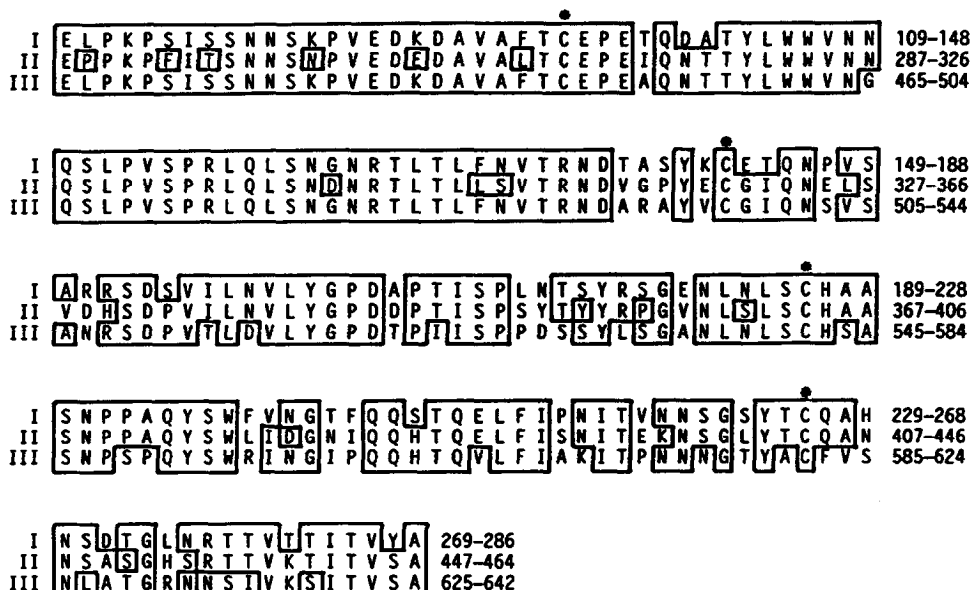


Fig.3. Comparison of amino acid sequences of homologous repetitive domains of the CEA. The three homologous domains are aligned for comparison of the amino acid sequences. The numbers of amino acid residues at the first and last ends of each line are given. Identical residues between domains are boxed. \* indicates cysteine residues.

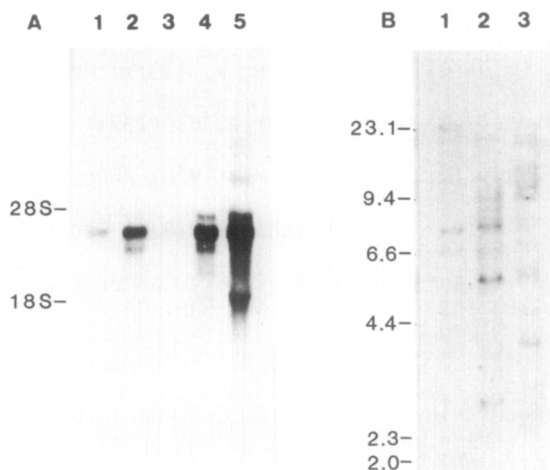


Fig.4. Autoradiograms of blot hybridization analysis of RNA from human tissues (A), and of restriction fragments of human genomic DNA (B). In A, poly(A)<sup>+</sup> RNA equivalent to 150µg of total RNA was analysed as described in METHODS. RNA was prepared from; lane 1, colonic cancer tissue from patient A; lane 2, human gastric carcinoma cell line (KATO III) producing CEA; lane 3, human hepatoma maintained in nude mice (total RNA, 15µg); lane 4, colonic cancer tissue from patient B; lane 5, healthy colonic mucosa of patient A; In B, human lymphocyte DNA digested with restriction endonuclease was analysed as described in METHODS. Restriction endonuclease used was; lane 1, BamHI; lane 2, EcoRI; lane 3, HindIII. Molecular weight markers are DNA digested with HindIII run through the same gel.

different 3' termini are cloned from the Okayama-Berg library (Fig.2). The relationships of the two clones and the individual RNA are not clear at this stage. Another feature to be noted is that the amounts of the RNAs expressed in the tissues and cells of different sources including cancerous and healthy colonic tissues of the same patient differ significantly. It is too early to interpret these data at this stage, for we do not know whether the hybridization probe is specific to CEA encoding sequence or recognizes other CEA-related antigen encoding sequences as well.

Human lymphocyte DNA analysed by blot hybridization after restriction endonuclease digestion gives many fragments which hybridize with short 5'-terminal region of the cDNA (Fig.4B), indicating the presence of either multiple copies of a gene homologous to the putative CEA cDNA or multiple CEA-related genes. The latter is more likely, for the presence of several similar genes are expected from the biochemical and immunochemical data of CEA and CEA-like antigens (11).

#### REFERENCES

1. Gold, P. & Freedman, S.O. (1965) *J. Exp. Med.* 121, 439-462.
2. Terry, W.D., Henkart, P.A., Coligan, J.E. & Todd, C.W. (1972) *J. Exp. Med.* 136, 200-204.
3. Kuroki, M., Shinoda, T., Takayasu, T., Koga, Y. & Matsuoka, Y. (1982) *Mol. Immunol.* 19, 399-406.
4. Krupcey, J., Gold, P. & Freedman, S.O. (1968) *J. Exp. Med.* 128, 387-398.
5. Von Kleist, S., Chavanel, G. & Burtin, P. (1972) *Proc. Natl. Acad. Sci. USA*, 69, 2492-2494.
6. Burtin, P., Chavanel, G. & Hirsch-Marie, H. (1973) *J. Immunol.* 111, 1926-1928.
7. Engvall, E., Shively, J.E. & Wrann, M. (1978) *Proc. Natl. Acad. Sci. USA*, 75, 1670-1674.
8. Kessler, M.J., Shively, J.E., Pritchard, D.G. & Todd, C.W. (1978) *Cancer Res.* 38, 1041-1048.
9. Kuroki, M., Koga, Y. & Matsuoka, Y. (1981) *Cancer Res.* 41, 713-720.
10. Svenberg, T. (1976) *Int. J. Cancer*, 17, 588-596.
11. Rogers, G.T. (1983) *Biochim. Biophys. Acta*, 695, 227-249.
12. Deininger, P., Jolly, D.J., Rubin, C.M., Friedmann, T. & Schmid, C.W. (1981) *J. Mol. Biol.* 151, 17-33.
13. Blobel, G. & Dobberstein, B. (1975) *J. Cell Biol.* 67, 835-851.
14. Perlman, D. & Halvorson, H.O. (1983) *J. Mol. Biol.* 167, 391-409.
15. Krantz, M.J. & Laferte, S. (1983) *Mol. Immunol.* 20, 409-420.
16. Oikawa, S., Inai, M., Ueno, A., Tanaka, S., Noguchi, T., Nakazato, H., Kangawa, K., Fukuda, A., & Matsuo, H. (1984) *Nature*, 309, 724-726.
17. Young, R.A. & Davis, R.W. (1983) *Proc. Natl. Acad. Sci. USA*, 80, 1194-1198.
18. Watson, C.J. & Jackson, J.F. (1985) In *DNA cloning* (D.M. Glover, ed.) Vol.1, pp. 79-88. IRL Press, Oxford, Washington, DC.
19. Okayama, H. & Berg, P. (1982) *Molec. Cell. Biol.* 2, 161-170.
20. Dale, R.M.K., McClure, B.A. & Hauchins, J.P. (1985) *Plasmid*, 13, 31-40.
21. Sanger, F., Nicklen, S. & Coulson, A.R. (1975) *Proc. Natl. Acad. Sci. USA*, 72, 3961-3965.
22. Vieira, J. & Messing, J. (1982) *Gene*, 19, 259-268.
23. Kyte, J. & Doolittle, R.F. (1982) *J. Mol. Biol.* 157, 105-132.
24. Nambu, J.R., Taussig, R., Mahon, A.C. & Scheller, R.H. (1983) *Cell*, 35, 47-56.
25. Southern, E.M. (1975) *J. Mol. Biol.* 98, 503-517.