Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 2965–2969, May 1987
Medical Sciences

Molecular cloning of a gene belonging to the carcinoembryonic antigen gene family and discussion of a domain model

(tumor antigens/sequence homologies/oligonucleotide probes/genomic sequence/exon structure)

JOHN A. THOMPSON*, HEMA PANDE, RAYMOND J. PAXTON, LOUISE SHIVELY, A. PADMA, ROBERT L. SIMMER†, CHARLES W. TODD, ARTHUR D. RIGGS, AND JOHN E. SHIVELY‡

Divisions of Immunology and Biology, Beckman Research Institute of the City of Hope, Duarte, CA 91010

Communicated by Rachmiel Levine, January 2, 1987 (received for review September 10, 1986)

ABSTRACT **Carcinoembryonic antigen (CEA) is a glycoprotein important as a tumor marker for colonic cancer. Immunological and biochemical studies have shown it to be closely related to a number of other glycoproteins, which together make up a gene family. We have cloned a member of this gene family by using long oligonucleotide probes (42–54 nucleotides) based on our protein sequence data for CEA and NCA (nonspecific cross-reacting antigen) and on human codon usage. The clone obtained (λ39.2) hybridizes with six probes and has a 15-kilobase insert. The 5′ end of the gene is contained within a 2700-base-pair *Eco*RI fragment, which hybridizes with five of the six synthetic probes. Sequencing of the 5′ end region revealed the location and structure of one exon and two putative intron boundaries. The exon encodes part of the leader sequence and the NH_2-terminal 107 amino acids of NCA. Southern blot analysis of human normal and tumor DNA, using as probes two λ39.2 fragments that contain coding sequences, suggests the existence of 9–11 genes for the CEA family. One of the restriction fragments described here has been used by Zimmermann *et al.* [Zimmermann, W., Ortlieb, B., Friedrich, R. & von Kleist, S. (1987) *Proc. Natl. Acad. Sci. USA* 84, 2960–2964] to isolate partial cDNA clones for CEA. The identity of this clone was verified with our protein sequence data [Paxton, R., Mooser, G., Pande, H., Lee, T. D. & Shively, J. E. (1987) *Proc. Natl. Acad. Sci. USA* 84, 920–924]. We discuss a domain structure for CEA based on the CEA sequence data and the NCA exon sequence data. It is likely that this gene family evolved from a common ancestor shared with neural cell adhesion molecule and α_1 B-glycoprotein and is perhaps a family within the immunoglobulin superfamily.**

Carcinoembryonic antigen (CEA), first discovered by Gold and Freedman (1), is a high molecular weight glycoprotein (M_r, 180,000), with considerable clinical utility (2–5) as a tumor marker for colonic cancer. Immunological and biochemical studies have revealed that CEA is a member of a family of closely related molecules, which share common antigenic determinants (6). CEA-related antigens, such as nonspecific cross-reacting antigen (NCA) and biliary glycoprotein, occur in normal tissues and may interfere with the determination of CEA in serum and tissue samples.

NCA (7) exists in at least two forms (8, 9): NCA-95 (M_r, 95,000), found in lung and spleen granulocytes; and NCA-55 (M_r, 55,000), found in both granulocytes and epithelial cells (9). A tumor-extracted antigen isolated in this laboratory (10) is similar to NCA-95. Although NH_2-terminal sequences have been analyzed on CEA, tumor-extracted antigen, and NCA (6), the presence of 40–60% carbohydrate severely hampered further structural studies until recently (11). Other cross-reacting antigens are NCA-2 (M_r, 160,000), isolated from meconium (12), and biliary glycoprotein (M_r, 85,000), isolated from bile (13). Recently, four CEA-like antigens (M_r, 200,000, 180,000, 114,000, and 85,000) have been described in human serum (14). The M_r 180,000 antigen corresponds to CEA and the M_r 85,000 antigen corresponds to biliary glycoprotein. CEA-like antigens have been described from colorectal carcinoma tissue (M_r, 128,000) and from meconium (M_r, 100,000) (15), as well as from human feces (16). Thus, it appears that at least seven distinct antigenic molecules exist within the CEA family.

Although the mRNA for CEA has been identified through *in vitro* translation and immunoprecipitation experiments (17), attempts to clone a CEA cDNA had been unsuccessful for many years in a number of laboratories, including our own. For this reason, we have used our protein sequence data for CEA, NCA-55, and tumor-extracted antigen (NCA-95) to synthesize long unique oligonucleotide probes for screening a human genomic library. We report here our initial studies on the isolation of a member of the CEA family. The NCA clone reported here also led directly to the identification of cDNA clones for CEA, as described in the companion paper by Zimmermann *et al.* (18).

MATERIALS AND METHODS

Bacterial Strains, Human Genomic Library, and Cloning Vectors. The λ Charon 4A human genomic library was from Lawn *et al.* (19) and was amplified in *Escherichia coli* strain LE392. Fragments were subcloned in the plasmid pUC18 and grown in TB1 bacteria.

Screening of the Human Genomic Library. Oligonucleotides were synthesized by the phosphoramidite procedure (20). The human genomic DNA library was screened by the method of Benton and Davis (21). Oligonucleotides were 5′-end-labeled (22) and hybridized as described (23). Washing was in 6× SSC/0.1% $NaDodSO_4$ (1× SSC = 0.15 M NaCl/0.015 M sodium citrate) three times, 15 min each, at room temperature, followed by 15-min washes at higher temperatures.

Restriction Analyses and Determination of DNA Sequences. Restriction enzyme analysis with *Eco*RI and *Sst* I was performed under standard conditions (22). Subcloned restriction fragments in phage M13mp18 M13mp19 were sequenced by the dideoxy-chain-termination method (24), using the universal primer or 17-mer synthetic oligonucleotides derived from known DNA sequence.

Extraction of Human DNA for Determination of the Size of the CEA Gene Family. Human DNA was extracted from 2–3

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

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

*Present address: Institut fur Immunbiologie, Stefan-Meier Str. 8, D7800 Freiburg, F.R.G.
†Present address: Abbott Laboratories, Abbott Park, Illinois.
‡To whom reprint requests should be addressed.

g of solid tissue (human colon tumor and normal human colon) or cell culture (LS174T) after solubilization in 1% $NaDodSO_4$/50 mM Tris·HCl pH 7.5/20 mM Na_2EDTA, containing proteinase K (100 μg/ml) at 37°C overnight. CsCl purification of the DNA was as described (25). After digestion to completion with excess restriction endonuclease and separation on a 0.8% agarose gel (10–20 μg of DNA per lane), the DNA was transferred to a nitrocellulose filter (26). Restriction fragments of the cloned gene were labeled by random priming (27) for hybridization with the filter (22).

RESULTS

Oligonucleotide Synthesis and Screening the Human Genomic Library. Based on known protein sequence data for CEA, NCA-55, and NCA-95 (tumor-extracted antigen) (11), six oligonucleotides were synthesized with unique sequences, utilizing optimal codon usage (28). Fig. 1 shows the nucleotide sequence of the probes; in each case, the complementary DNA strand was synthesized. The optimal hybridization conditions for screening the human genomic library were based on formulae developed by Lathe (28), which take into account oligonucleotide length and G + C base composition, and which assume an 85% sequence homology.

Oligonucleotide 3 was used for the initial screening of ≈500,000 recombinant phages. This sequence corresponds to the known NH_2-terminal region, which is common to CEA and NCA (11). From 79 plaques, which were positive when hybridized at a relatively low stringency (35°C in 1 M salt/20% formamide and washed up to 50°C for 15 min in 1 M salt/0.1% $NaDodSO_4$), 69 could be recognized on the rescreen with the same oligonucleotide at the same stringency. All 69 were washed in temperature increments of 5°C, using the same buffer concentrations, until all the radioactivity had disappeared (up to 70°C wash for 15 min).

Using the formulae given by Lathe (28), 14 clones showed an estimated sequence homology of >80% (still visible after a 60°C wash), and one clone (λ39.2) showed an estimated sequence homology of >85% (still visible after a 65°C wash). The stripped filters were rehybridized with an equal mixture of oligonucleotides CEA 3+ and NCA 3+, both of which had comparable theoretical melting temperatures. These oligonucleotides were synthesized based on protein sequence data for the protein segments directly adjacent to the common

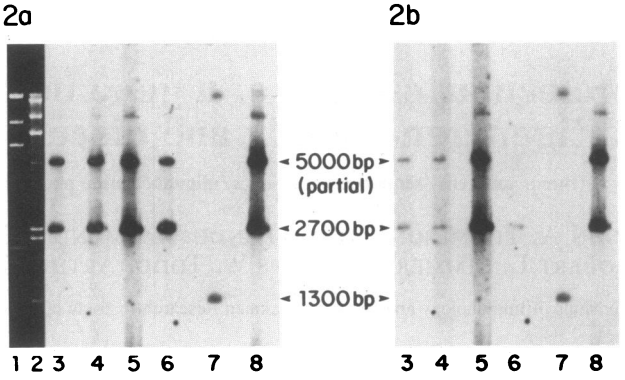

FIG. 2. Blot analysis of clone λ39.2 DNA. The clone was hybridized with six synthetic oligonucleotides (see Fig. 1). Lane 1, length standard, λ *Hin*dIII; lanes 2–8, *Eco*RI-digested λ39.2 DNA; lane 2, stained with ethidium bromide; lanes 3–8, hybridized with oligonucleotide probes 3, CEA 3+, NCA 3+, 1, 2, and 4, respectively. (*a*) Lower stringency wash (60°C, 15 min in 6× SSC/0.1% $NaDodSO_4$ for all except probe 2, which was a 55°C wash). (*b*) Higher stringency wash (65°C, 15 min in 6× SSC/0.1% $NaDodSO_4$ for all except probe 2, which was a 60°C wash).

NH_2-terminal oligonucleotide 3 sequence. In this region, CEA and NCA show some sequence divergence as shown in Fig. 1. From the 69 clones that were positive when rescreened with oligonucleotide 3, only one clone hybridized with the CEA 3+/NCA 3+ mixture, even at a relatively low stringency (42°C hybridization; 55°C wash). This clone (λ39.2) still hybridized at a high-stringency wash (65°C wash for 15 min), showing an estimated sequence homology of >81% with oligonucleotides CEA 3+ and/or NCA 3+.

Fig. 2*a* shows a Southern blot analysis of λ39.2 DNA. All six oligonucleotides hybridize with the insert, each showing >80% estimated sequence homology. Five of the six oligonucleotides, including those from the NH_2 terminus, hybridize with one *Eco*RI fragment [2700 base pairs (bp)]. Only oligonucleotide 2 hybridizes to a different *Eco*RI fragment (1300 bp). Fig. 2*b* shows the same filters washed at higher stringencies. It is obvious that NCA 3+ still remains bound, whereas CEA 3+ has virtually disappeared, although both have very similar theoretical melting temperatures. This indicates that the gene is more similar to NCA than to CEA.

Restriction Map of Clone λ39.2 and Sequence Analysis of

PROBE		
1	PROTEIN SEQUENCE	Q I I G Y V I G T Q Q A T P G P A Y
	CODING STRAND	5'-CAGATTATTGGCTATGTGATTGGCACCCAGCAGGCTACCCCAGGCCCAGCTTAC-3'
	COMPLEMENTARY STRAND	3'-GTCTAATAACCGATACACTAACCGTGGGTCGTCCGATGGGGTCCGGGTCGAATG-5'
	NCA EXON	5'-CTAATTGTAGGATATGTAATAGGAACTCAACAAGCTACCCCAGGGCCCGCATAC-3'
2	PROTEIN SEQUENCE	N D A G S Y E C E I Q N P A
	CODING STRAND	5'-AATGATGCTGGCTCCTATGAATGTGAAATTCAGAACCCAGCC-3'
	COMPLEMENTARY STRAND	3'-TTACTACGACCGAGGATACTTACACTTTAAGTCTTGGGTCGG-5'
3	PROTEIN SEQUENCE	T I E S T P F N V A E G K E V
	CODING STRAND	5'-ACCATTGAATCCACCCCATTCAATGTGGCTGAAGGCAAGGAAGTG-3'
	COMPLEMENTARY STRAND	3'-TGGTAACTTAGGTGGGGTAAGTTACACCGACTTCCGTTCCTTCAC-5'
	NCA EXON	5'-ACTATTGAATCCACGCCATTCAATGTCGCAGAGGGGAAGGAGGTT-3'
CEA 3+	PROTEIN SEQUENCE	V H N L P Q H L F G Y S W Y K
	CODING SEQUENCE	5'-GTGCACAACCTGCCCCAGCACCTGTTTGGCTACTCCTGGTACAAG-3'
	COMPLEMENTARY STRAND	3'-CACGTGTTGGACGGGGTCGTGGACAAACCGATGAGGACCATGTTC-5'
NCA 3+	PROTEIN SEQUENCE	A H N L P Q N R I G Y S W Y K
	CODING STRAND	5'-GCCCACAACCTGCCCCAGAACCGGATTGGCTACTCCTGGTACAAG-3'
	COMPLEMENTARY STRAND	3'-CGGGTGTTGGACGGGGTCTTGGCCTAACCGATGAGGACCATGTTC-5'
	NCA EXON	5'-GCCCACAACCTGCCCCAGAATCGTATTGGTTACAGCTGGTACAAA-3'
4	PROTEIN SEQUENCE	V N E E A T G Q F H V Y P E L P K P
	CODING SEQUENCE	5'-GTGAATGAAGAAGCCACTGGCCAGTTCCATGTGTATCCAGAACTGCCAAAGCCA-3'
	COMPLEMENTARY STRAND	3'-CACTTACTTCTTCGGTGACCGGTCAAGGTACACATAGGTCTTGACGGTTTCGGT-5'
	NCA EXON	5'-GTGAATGAAGAAGCAACCGGACAGTTCCATGTATAC-3'

FIG. 1. Oligonucleotide probes for CEA and cross-reacting antigens. The synthetic oligonucleotides with unique sequences are based on protein sequence data for CEA and cross-reacting antigens NCA-55 and NCA-95 (11). In each case, the complementary strands were synthesized. The positions of these probes in the cloned NCA exon are shown in Fig. 4. The exon sequences found in clone λ39.2 are shown for probes 1, 3, 3+, and 4. Amino acids are designated by the single-letter code.

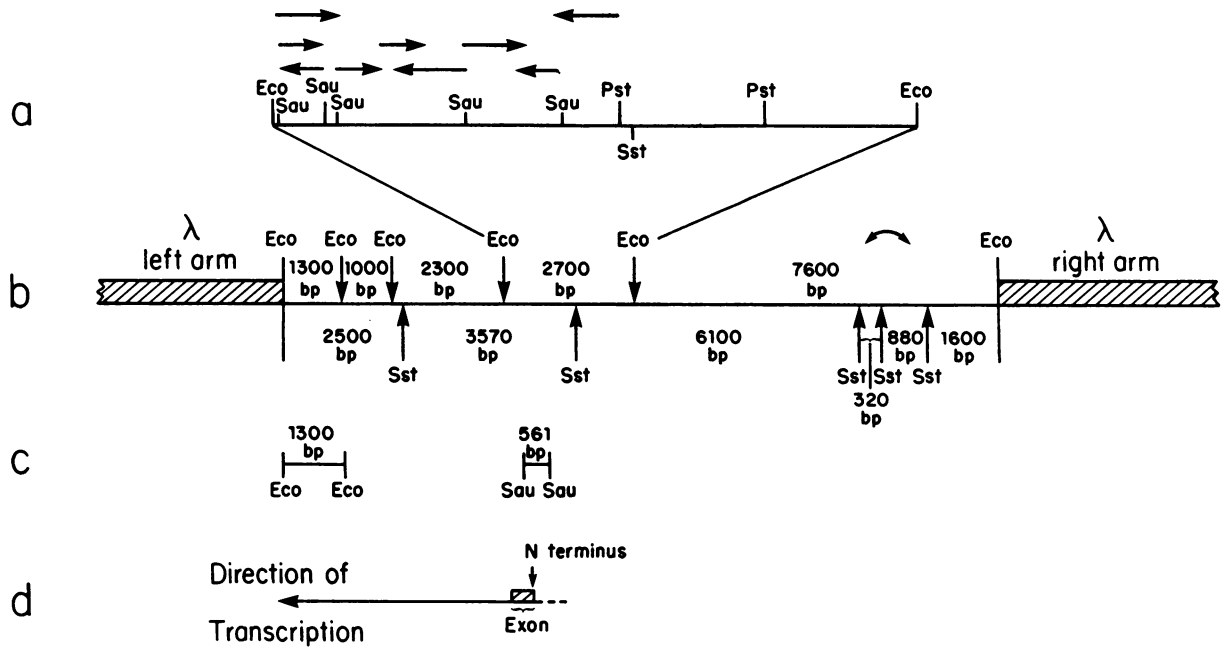

FIG. 3. Restriction map and sequencing strategy of the NH_2-terminal region of the DNA insert of clone λ39.2. (*a*) Restriction map and sequencing strategy of the 2700-bp *Eco*RI fragment shown in *b*. The *Sau*3a positions outside of the sequenced region have not been determined. (*b*) Restriction map of λ39.2. There is a small uncertainty upstream from the NH_2-terminal region of the gene. (*c*) Restriction fragments used to probe for other members of the CEA gene family. Both contain coding information (see Fig. 2). (*d*) Direction of transcription of the gene, including the position of the exon shown in Fig. 4. The NH_2 terminus of the coding sequence for the mature protein (minus leader sequence) is indicated.

the NH_2-Terminal Region of the Gene. Fig. 3 shows the *Eco*RI and *Sst* I restriction map for clone λ39.2 (Fig. 3*b*). The orientation of the gene (Fig. 3*d*) can be determined, since we know that the 1300-bp *Eco*RI fragment that binds oligonucleotide 2 must be downstream from the 2700-bp *Eco*RI fragment containing the NH_2 terminus. The orientation of the 1300- and 1000-bp *Eco*RI fragments with respect to each other was determined through partial digestion of the largest *Sst* I fragment with *Eco*RI, which showed the 1000-bp *Eco*RI fragment to be adjacent to the 2300-bp *Eco*RI fragment (data not shown).

Fig. 4 shows the DNA sequence for the *Pst* I/*Eco*RI fragment (see Fig. 3*a*). A region near the 3′ end of this fragment contains DNA sequences corresponding to four of the six oligonucleotide probes. In Fig. 5, a comparison of the protein sequence translated from this exon with known protein sequence data for CEA and NCA (11) identifies this gene as encoding NCA. Putative intron boundaries at each

```
        10        20        30        40        50        60        70        80        90       100
         +         +         +         +         +         +         +         +         +         +
CTGCAGATTGCATGTCCCCTGGAAGGAGGTCCTGCTCACAGGTGAGGGGAGGACTCCCTCGGAGTGGATGGGAGGAGGGAGCACAGAGACTGGCTAGGGT  100

CTCCTGGGGAGGACAAGGCTCTGAGAGGAGACAGAGGGCTTTTGTTGAAGCCTGAGGAAACAGAACACCAGAGAGGGACAGGGGTCACAACAGGAAAGTC  200

ACACTAAACTGGGATTGATAAAAAGGGAGGAAAATCAATTGATCATGTTTTCCAAGTTAATCATCATTTGTCATTACCATTTGAAAAAAAAGAAAAATGA  300

TAGAAATCAGAACTGCATTAGGATGACACTCCAAATAAAAATATAACAAGGAAACTAAATGCTGCCCTTACTCACCAATCAGAAGTTGAAAAATAACCAC  400

CAGATACACTCATTAACTCATCCACAAGCATTTGCAATCAATTTTAGTCAATGGCATACAACAAGCATCAGACAAGTCTCAGTCATCACAGAGCTTATGC  500

TGTCATGAAGAGGAAAACACACACACAAAGAGATATAGAATGTGAGGTCAGGTGTTGACAAGAGCCCTGGAAGGAACAGAGCAGGGAAAGGTCAGAAAGA  600

AAAGACCCAGGGTCTGTAGAGGGGGTGTCAGGGAAGGGATCTCCCAAGAATGCCCTGATGTGAGCAGGACCTGAGGCCAGTGGGGAGGGAGCCATGCAGA  700

CCCCTGGGGAAGAGCATTCCACACAGGGAAATGCCAAGGTCAAAGGTGCTGAAGGAATGGGGGTGTCACACTGCTGACTTTGACTCAGTAGGACACACAC  800

ACACACACACACACACACACACACACGCTCCAACGTGGAGGGGTGAAGAGACCTGCTCAGGACCCAGGGCCCTGTTTTTCCACCCTAATGCATAGGTC  900

CCAATATTGACCGATGCTCTCTCCTCTCTCCTAGCCTCACTTCTAACCTTCTGGAACCCACCCACCACTGCCAAGCTCACTATTGAATCCACGCCATTCA 1000
                                SerLeuLeuThrPheTrpAsnProProThrThrAlaLysLeuThrIleGluSerThrProPheA
                                /----------Leader peptide---------->

ATGTCGCAGAGGGGAAGGAGGTTCTTCTACTCGCCCACAACCTGCCCCAGAATCGTATTGGTTACAGCTGGTACAAAGGCGAAAGAGTGGATGGCAACAG 1100
snValAlaGluGlyLysGluValLeuLeuLeuAlaHisAsnLeuProGlnAsnArgIleGlyTyrSerTrpTyrLysGlyGluArgValAspGlyAsnSe

TCTAATTGTAGGATATGTAATAGGAACTCAACAAGCTACCCCAGGGCCCGCATACAGTGGTCGAGAGACAATATACCCCAATGCATCCCTGCTGATCCAG 1200
rLeuIleValGlyTyrValIleGlyTyrGlnGlnAlaThrProGlyProAlaTyrSerGlyArgGluThrIleTyrProAsnAlaSerLeuLeuIleGln

AACGTCACCCAGAATGACACAGGATTCTATACCCTACAAGTCATAAAGTCAGATCTTGTGAATGAAGAAGCAACCGGACAGTTCCATGTATACCGTGAGT 1300
AsnValThrGlnAsnAspThrGlyPheTyrThrLeuGlnValIleLysSerAspLeuValAsnGluGluAlaThrGlyGlnPheHisValTyr

ATTTCCACATGACCTCTGGGTGTTGGGGGTCAGTTCTACTTCCCACATACGGGATTGTCAGGCCTGGGTTGTGCCTGTGGCCCTCTCTGCATTACATCCT 1400

GTATCAGGGTTTGGACATTTAGTGCAGGACACACACGGGGAAGACAAACTTCCACAGATCAGAATTC
```

FIG. 4. DNA sequence data for the 1467-bp *Eco*RI/*Pst* I fragment shown in Fig. 3*a*. Numbering of nucleotides is based on their position with regard to the *Pst* I restriction site. The positions of probes 1 (nucleotides 1102–1155), 3 (979–1023), 3+ (1033–1077), and 4 (1258–1293) are shown with overbars. The boxed area indicates a potential regulatory sequence.

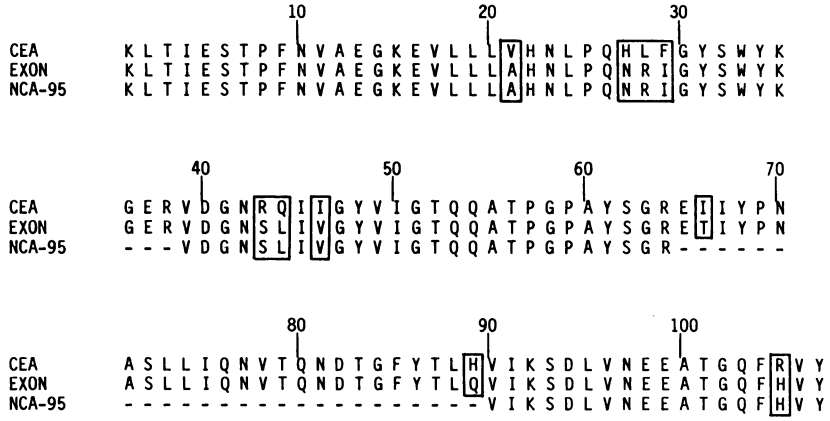

FIG. 5. Comparison of known protein sequences for CEA, NCA-55, and NCA-95 (11), with the exon translated from clone λ39.2. Boxed areas indicate amino acid differences. Dashes indicate gaps in amino acid sequence data. Amino acids are designated by the single-letter code.

end of this region were recognized through homologies to consensus sequences for acceptor and donor sites (29). This suggests that the *Pst* I/*Eco*RI fragment contains an exon that encodes the NH_2-terminal 107 amino acids of NCA and part of the leader sequence (see *Discussion*).

Estimation of the Size of the CEA Gene Family. Two restriction endonuclease fragments from clone λ39.2, which had been shown to contain coding sequence through their hybridization to oligonucleotide 2 (*Eco*RI 1300-bp fragment) or through sequencing the NH_2-terminal region (*Sau*3a 561-bp fragment), were independently used to probe digests of total human DNA. Fig. 6 shows the Southern blot analysis. Up to 11 bands are seen at lower stringencies. This number is comparable for both probes. No significant difference is seen in the hybridization patterns between DNA from normal and from tumor tissues. Lane 4 shows a much lower hybridization due to the small amount of DNA loaded on the gel. When the filters were washed at a higher stringency, only one fragment remained for each probe (indicated with arrows in Fig. 6), with a size expected from clone λ39.2.

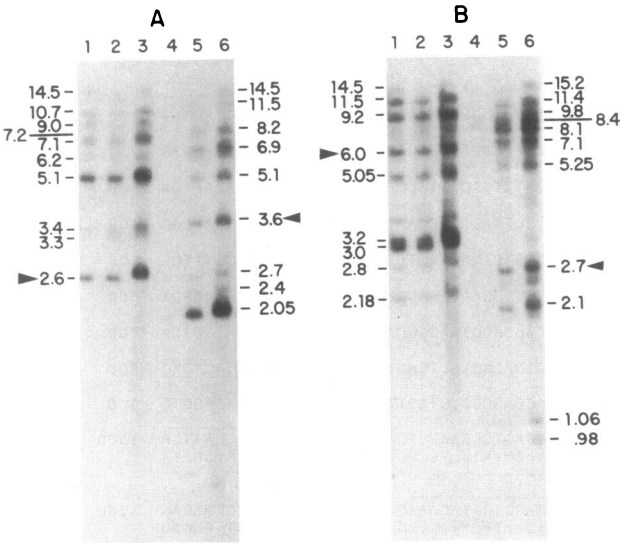

FIG. 6. Blot analysis pattern of total human DNA digested with two restriction fragments isolated from clone λ39.2 (see Fig. 3*c*). (*A*) Hybridization with the 561-bp *Sau*3a fragment. (*B*) Hybridization with the 1300-bp *Eco*RI fragment. Lanes 1 and 4, DNA from normal human leukocytes; lanes 2 and 5, DNA from human colon tumor; lanes 3 and 6, DNA from human colon tumor cell line (LS174T); lanes 1–3, *Eco*RI digest of total human DNA; lanes 4–6, *Sst* I digest of total human DNA. Filters were washed at 65°C in 0.5× SSC/0.5% $NaDodSO_4$. Arrows indicate the fragments remaining at a high-stringency wash (68°C in 0.1× SSC/0.5% $NaDodSO_4$).

DISCUSSION

Using long synthetic oligonucleotides with unique sequences to screen a human genomic library, we have been able to isolate a gene belonging to the CEA family. Comparative hybridization with probes for NCA or CEA allowed provisional identification of the gene as NCA. This initial assignment was confirmed through sequence analysis of the NH_2-terminal region (Fig. 4). A comparison of the exon-translated sequence and the known protein sequence for NCA showed them to be identical (Fig. 5).

A close homology can also be noted between CEA and NCA (89%), which indicates either conservation of protein sequence or a recent evolutionary divergence of the two genes. It is too soon to be able to identify the gene as encoding NCA-55, NCA-95, or yet another NCA gene.

The putative exon boundaries were identified through sequence homologies to a consensus of intron donor and acceptor sequences (29). Further evidence to support the presence of an intron boundary at the 3′ end of this exon was deduced from a known protein sequence in this region. Oligonucleotide 4 (Fig. 1), which is based on a known peptide sequence, is apparently situated at the 3′ boundary of the exon and is interrupted by the presence of the intron. The DNA sequence diverges from the oligonucleotide sequence at this point, indicating the presence of an intron boundary as indicated in Fig. 4. The continuation of the oligonucleotide 4 sequence is expected further downstream.

The putative acceptor sequence at the 5′ end of this exon is in the leader region, for which no known protein sequence data are available. However, an in-frame continuation of the sequence upstream would lead to a stop codon before an initiation codon is reached. Because of the strong sequence homology with the acceptor consensus sequence (29) at the point indicated in Fig. 4, we are predicting an intron/exon boundary.

Immunological and biochemical studies have indicated the existence of at least seven different proteins in the CEA family. The results of our Southern analysis of genomic DNA (Fig. 6) further support the idea of a CEA gene family. Two cloned fragments (561-bp *Sau*3a and 1300-bp *Eco*RI) were used as probes, and 9–11 bands are seen. The partial cDNA sequence for CEA shows the existence of repeating units with high internal sequence homologies (18). The exon shown here for NCA, which contains the coding sequence for the NH_2-terminal portion of the protein, exhibits lower homology to the repeat units than the repeating units show to themselves. Our data suggest that the NH_2-terminal region is also shared by the other members of the CEA family.

The 1300-bp *Eco*RI restriction fragment and synthetic oligonucleotides (Figs. 1 and 3) were used by Zimmermann *et*

al. (18) to identify two CEA cDNA clones in a colon tumor cDNA library. Sequence analysis of these cDNAs showed internal homology at the level of 1 repeat per 4 cysteine residues. A total of 3 repeats (12 cysteines) is predicted for CEA from protein sequence data (11). All of the cysteines in CEA are known to be in disulfide bonds, resulting in a total of six disulfide bonds. The presence of internal repeats within the CEA protein suggests that the CEA gene evolved from a more primitive gene encoding a maximum of 4 cysteine residues (two disulfide loops). The amino acid spacing between the cysteine residues ranges from 39 to 47 residues (18). The NH_2-terminal domain is missing the cysteine residues but exhibits some sequence homology to the disulfide loops, suggesting a special function for the NH_2-terminal domain. The NH_2-terminal exon for NCA reported in this paper indicates the precise boundary between the end of the cysteine-lacking region and the beginning of the first disulfide loop repeat. These data suggest that exons in CEA and NCA may correspond to homologous domains. Fig. 7 presents a domain model of CEA based on information for CEA at the protein and nucleotide level (11, 18), the NCA exon data presented here, and the fact that CEA is a member of the immunoglobulin gene superfamily (11). This model explains many of the known structural and immunological features of the CEA and CEA-like antigens—for example, the majority of the epitopes recognized by antibodies to CEA are conformationally dependent, and their antibody-binding ability is lost upon reduction and S-alkylation of cysteine residues (6). It is reasonable to predict that these epitopes may reside in the disulfide loops. The internal repeats account for the ability of one monoclonal antibody to bind two or more times to one CEA molecule (30). Epitopes located within the NH_2-terminal (and possibly the COOH-terminal) domain would be less likely to be conformationally labile, thus accounting for additional epitopes reported in CEA (31).

This model has striking resemblances to the structure of neural cell adhesion molecule (32) and α_1B-glycoprotein (33). These proteins have five disulfide loops, comprising adjacent cysteine residues with a spacing of 43–49 residues between cysteines and possessing high internal sequence homology. The sequences near the loop-forming cysteine residues of CEA are homologous to the sequences near the cysteine residues in the immunoglobulin superfamily (11).

The CEA family also provides an interesting system for gene expression studies. Although different members of the family show strong sequence homology—e.g., CEA and NCA (Fig. 5)—they are expressed in different tissues and at different times. Therefore, a comparison of regulatory regions for these different genes may help identify tissue-specific control sequences. A potential regulatory region is visible in Fig. 4 from nucleotides 793 to 828 (boxed area). This poly(dCA) region has been reported in a number of genes (34). For NCA, it is situated in the putative first intron, ≈100 nucleotides upstream from the exon acceptor site.

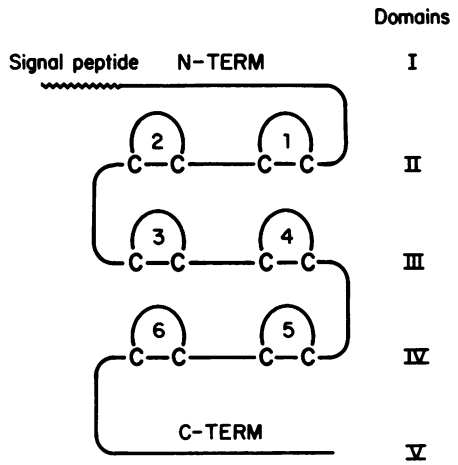

FIG. 7. Domain model for CEA. Six disulfide loops with internal homologies for loops 1, 3, 5 and 2, 4, 6 are shown. The NH_2-terminal domain exhibits some homology to domains II, III, and IV. The short COOH-terminal domain shows no homology to the other domains (18). It is probably a membrane attachment segment.

This article is dedicated to the memory of Charles W. Todd, who contributed much to the field of cancer research and who died January 1, 1987. The authors gratefully acknowledge discussions with Dr. Nozomu Mori regarding homology to the immunoglobulin superfamily. This research was supported by National Cancer Institute Grant CA37808. This work was made possible in part by support from the Sanford Sigoloff Research Fellowship.

1. Gold, P. & Freedman, S. O. (1965) *J. Exp. Med.* **121,** 439–462.
2. Fritsche, R. & Mach, J.-P. (1977) *Immunochemistry* **14,** 119–127.
3. Egan, M. L., Pritchard, D. G., Todd, C. W. & Go, V. L. W. (1977) *Cancer Res.* **37,** 2638–2643.
4. 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.
5. Mackay, I. R. (1979) in *Immunodiagnosis of Cancer,* eds. Herberman, R. B. & McIntire, K. R. (Dekker, New York), Part I, pp. 255–264.
6. Shively, J. E. & Todd, C. W. (1980) in *Cancer Markers: Diagnostic and Developmental Significance,* ed. Sell, S. (Humana, Clifton, NJ), pp. 295–314.
7. von Kleist, S., Chavanel, G. & Burtin, P. (1972) *Proc. Natl. Acad. Sci. USA* **69,** 2492–2494.
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,** 643–649.
10. Kessler, M. J., Shively, J. E., Pritchard, D. G. & Todd, C. W. (1978) *Cancer Res.* **38,** 1041–1048.
11. Paxton, R., Mooser, G., Pande, H., Lee, T. D. & Shively, J. E. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 920–924.
12. Burtin, P., Chavanel, G. & Hirsch-Marie, H. (1973) *J. Immunol.* **111,** 1926–1928.
13. Svenberg, T. (1976) *Int. J. Cancer* **17,** 588–596.
14. Neumaier, M., Fenger, U. & Wagener, C. (1985) *Mol. Immunol.* **22,** 1273–1277.
15. Neumaier, M., Fenger, U. & Wagener, C. (1985) *J. Immunol.* **135,** 3604–3609.
16. Kuroki, M., Koga, Y. & Matsuoka, Y. (1981) *Cancer Res.* **41,** 713–720.
17. Zimmermann, W., Friedrich, R., Grunert, F., Luckenbach, G.-A., Thompson, J. A. & von Kleist, S. (1983) *Ann. N.Y. Acad. Sci.* **417,** 21–30.
18. Zimmermann, W., Ortlieb, B., Friedrich, R. & von Kleist, S. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 2960–2964.
19. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. & Maniatis, T. (1978) *Cell* **15,** 1157–1174.
20. Caruthers, M. H., Beaucage, S. L., Efcavitch, J. W., Fisher, E. F., Goldman, R. A., Dehaseth, P. L., Mandecki, W., Naheucci, M. S., Rosendahl, M. S. & Stabinsky, Y. (1982) *Cold Spring Harbor Symp. Quant. Biol.* **47,** 411–418.
21. Benton, W. D. & Davies, R. W. (1977) *Science* **196,** 180–182.
22. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
23. Ullrich, A., Berman, C. M., Dull, T. J., Gray, A. & Lee, J. M. (1984) *EMBO J.* **3,** 361–364.
24. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
25. Flamm, W. G., Birnstiel, M. L. & Walker, P. M. B. (1972) in *Subcellular Components: Preparation and Fractionation,* ed. Birnie, G. D. (Butterworth, London), pp. 293–294.
26. Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517.
27. Feinberg, A. P. & Vogelstein, B. (1983) *Anal. Biochem.* **132,** 6–13.
28. Lathe, R. (1985) *J. Mol. Biol.* **183,** 1–12.
29. Mount, S. M. (1982) *Nucleic Acids Res.* **10,** 459–472.
30. Hedin, A., Hammarstrom, S. & Larsson, A. (1982) *Mol. Immunol.* **19,** 164–168.
31. Haggarly, A., Legler, C., Krantz, M. J. & Fuks, A. (1986) *Cancer Res.* **46,** 300–309.
32. Hemperly, J. J., Murray, B. A., Edelman, G. M. & Cunningham, B. A. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 3037–3041.
33. Ishioka, N., Takahashi, N. & Putnam, F. W. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 2363–2364.
34. Hamada, H., Petrino, M. G., Kakunaga, T., Seidman, M. & Stollar, B. D. (1984) *Mol. Cell. Biol.* **4,** 2610–2621.