Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 6959–6963, September 1988
Medical Sciences

Molecular cloning of a cDNA coding biliary glycoprotein I: Primary structure of a glycoprotein immunologically crossreactive with carcinoembryonic antigen

(human colon cDNA library/nonspecific crossreacting antigen/immunoglobulin gene superfamily/human bile)

YUJI HINODA*, MICHAEL NEUMAIER*, STANLEY A. HEFTA*, ZOFIA DRZENIEK†, CHRISTOPH WAGENER†, LOUISE SHIVELY*, LAURA J. F. HEFTA*, JOHN E. SHIVELY*‡, AND RAYMOND J. PAXTON*

*Division of Immunology, Beckman Research Institute of the City of Hope, Duarte, CA 91010; and †Department for Clinical Chemistry and Pathobiochemistry, Medical Faculty, University of Technology, Pauwelsstrasse, D-5100 Aachen, Federal Republic of Germany

Communicated by Rachmiel Levine, May 23, 1988 (received for review April 12, 1988)

ABSTRACT **We have isolated and sequenced four overlapping cDNA clones from a normal adult human colon library, which together gave the entire nucleotide sequence for biliary glycoprotein I (BGP I). BGP I is a member of the carcinoembryonic antigen (CEA) gene family, which is a subfamily in the immunoglobulin gene superfamily. The deduced amino acid sequence of the combined clones for BGP I revealed a 34-residue leader sequence followed by a 108-residue N-terminal domain, a 178-residue immunoglobulin-like domain, a 108-residue region specific to BGP I, a 24-residue transmembrane domain, and a 35-residue cytoplasmic domain. The nucleotide sequence of BGP I exhibited greater than 80% identity with CEA and nonspecific crossreacting antigen (NCA) in the leader peptide, N-terminal domain, and immunoglobulin-like domain. The BGP I-specific domain, designated A′, was 56.7% and 55.8% identical at the nucleotide level and 42.6% and 39.6% identical at the amino acid level to the immunoglobulin-like domain of NCA and the first immunoglobulin-like domain of CEA, respectively. Beyond nucleotide position 1375 the 3′ region of the BGP I cDNA was found to be specific for BGP I. Hybridization of a probe from this region to electrophoretic blots of RNAs from different human tissues showed a predominant 2.8-kilobase (kb) message accompanied by weaker bands 4.1 and 2.1 kb in size. The same probe gave a single band in Southern blot analysis of restricted total human DNA. Using a coding region probe from the BGP I domain A′, we observed 4.1- and 2.1-kb messages. Lack of the 2.8-kb band suggested that different forms of BGP I may be generated by posttranscriptional modification of the same gene. We propose that BGP I diverged from NCA by acquiring an immunoglobulin-like domain substantially different from the domains found in NCA or CEA and also a new cytoplasmic domain. The latter feature should result in a substantially different membrane anchorage mechanism of BGP I compared to CEA, which lacks the cytoplasmic domain and is anchored via a phosphatidylinositol-glycan structure. Protein structural analysis of BGP I isolated from human bile revealed a blocked N terminus, 129 amino acids of internal sequence that are in agreement with the translated cDNA sequence, and five glycosylation sites in the peptides sequenced.**

Biliary glycoprotein I (BGP I), an antigen crossreactive with the carcinoembryonic antigen (CEA), was originally detected in normal human bile by using polyclonal anti-CEA antibodies (1) and can also be defined by crossreactivity with monoclonal anti-CEA antibodies with various epitope specificities (2). The molecular weight of BGP I is 85,000, and the antigen consists of a single polypeptide chain containing approximately 40% carbohydrate by weight (3). Analysis of the primary structure of the protein has been hampered by a blocked N terminus (unpublished data). While it is present in normal gallbladder or hepatic bile, little or no BGP I is found in the bile in diseases concomitant with the obstruction or inflammation of the biliary duct system (4). Instead, two other forms of BGP I are expressed, BGP II and BGP III, which have been partially characterized in immunological studies and appear to be closely related to nonspecific crossreacting antigen (NCA) and CEA, respectively (1). Increased serum levels of BGP I are found in individuals suffering from hepatic disorders. BGP I-specific antisera often show partial identity between serum BGP I and BGP I from normal bile, but no data are reported so far as to whether this immunological difference is caused by proteolysis or by the expression of different forms of BGP I in those diseases.

We report here the isolation and characterization of four overlapping clones obtained from an adult human colon library that together gave the entire coding sequence for BGP I as well as parts of the untranslated regions.§ Protein sequence analysis of BGP I isolated from normal human bile agreed with the translated cDNA sequence for the peptides thus far characterized. We discuss the evolutionary relationships among BGP I, NCA, and CEA.

MATERIALS AND METHODS

Chemicals. Restriction enzymes, the Klenow fragment of *Escherichia coli* DNA polymerase I, T4 polynucleotide kinase, and T4 DNA ligase were purchased from Bethesda Research Laboratories, Boehringer Mannheim, or New England Biolabs. Deoxynucleotides were obtained from Pharmacia. Radioisotopes were purchased from DuPont–New England Nuclear.

Library Screening and DNA Sequencing. A cDNA library constructed in λgt11 from adult human colon tissue was purchased from Clontech Laboratories (Palo Alto, CA). The library was screened according to Davis *et al.* (5), using DNA restriction endonuclease fragments labeled by random priming (6) or synthetic oligonucleotides labeled with T4 polynucleotide kinase (ref. 7, pp. 122–123). Positive clones were subcloned in Bluescript M13 KS(+) (Stratagene, San Diego, CA) and sequenced by the dideoxy chain termination method (8), using a Sequenase kit (United States Biochemical, Cleveland) as described (9).

RNA Blot Analysis. Total RNA was prepared from different tissues according to Meese and Blin (10). The following

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: BGP I, biliary glycoprotein I; NCA, nonspecific crossreacting antigen; CEA, carcinoembryonic antigen.

‡To whom reprint requests should be addressed.

§The sequence reported in this paper is being deposited in the EMBL/GenBank data base (IntelliGenetics, Mountain View, CA, and Eur. Mol. Biol. Lab., Heidelberg (accession no. J03858).

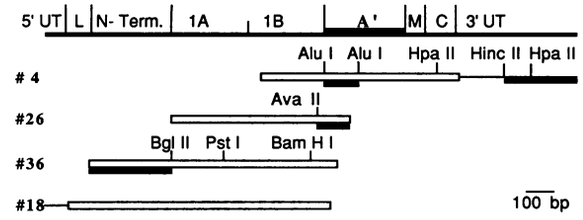

FIG. 1. The cDNA sequence of BGP I was determined in both directions from the four overlapping clones by using the restriction sites shown. The open boxes represent the open reading frame. BGP I-specific probes are given as closed boxes. The domain structure of BGP I is illustrated at the top of the figure with the BGP I-specific domain A' outlined. UT, untranslated region; L, leader peptide; 1A and 1B, subdomains of the immunoglobulin-like domain; M, membrane-spanning domain; C, cytoplasmic domain.

sources were used: human placenta, liver metastasis of a human colon carcinoma, leukemia cell line K562 [American Type Culture Collection (ATCC) CTL243], and mammary carcinoma cell line MCF7 (ATCC HTB22). Poly(A)$^+$ RNA from normal human liver was also used (gift from Stan Gartler). The RNA was electrophoresed under denaturing conditions (11) with recirculation of the electrophoresis buffer and blotted onto Gentrans 45 membrane (Plasco, Woburn, MA) by capillary transfer in 20× SSC buffer (1× SSC is 0.15 M NaCl/0.015 M sodium citrate, pH 7). After prehybridization, membranes were hybridized overnight (11) to radiolabeled probes from the specific domain in the coding region and the 3' untranslated region of the BGP I gene. Filters were washed at a final stringency of 0.1× SSC/0.1% sodium dodecyl sulfate at 65°C for 1 hr and exposed to x-ray film at −70°C with two intensifying screens.

Southern Blot Analysis. Human leukocyte DNA was prepared according to Kan *et al.* (12). The DNA was digested with restriction endonucleases *Eco*RI, *Sst* I, *Eco*RI plus *Sst* I, *Pst* I, *Bam*HI, *Pvu* II, and *Hin*dIII. The digested DNA (15 μg per lane) was electrophoresed on a 0.6% agarose gel and blotted onto nitrocellulose according to Wahl *et al.* (13). Hybridization to a ^{32}P-labeled 285-base-pair (bp) *Hinc*II/*Eco*RI fragment from the BGP I 3'-untranslated region was performed according to Maniatis *et al.* (ref. 7, pp. 387–389).

BGP I Isolation and Protein Sequence Analysis. BGP I was purified from normal human bile by immunoaffinity chromatography as previously described (2), followed by gel permeation and reverse-phase HPLC chromatography. Purified BGP I (40 μg) was deglycosylated, reduced, alkylated, and digested by chymotrypsin (14), and the resulting chymotryptic peptides were characterized by microsequence and fast atom bombardment–mass spectrometry analyses.

RESULTS

Characterization of cDNA Clones. A normal human colon cDNA library containing approximately 1 × 10^6 independent clones was screened with a 410-bp *Pst* I/*Pvu* II fragment from CEA cDNA coding for a portion of the second and third immunoglobulin-like domains of CEA (15). Sixty-one positive clones were rescreened with a 31-base synthetic oligonucleotide probe corresponding to nucleotides 60–90 of the N-terminal region of CEA (16). Three positives were subcloned and sequenced, and one of them (clone 36) was found to contain an insert coding for the N-terminal domain and the immunoglobulin-like domain of BGP I (see Fig. 1). Fifty-eight

```
          GTGACAGAGCAGCCGTGCTCGAAGCGTTCCTGGAGCCCAAGCTCTCCTCCACAGGTGAAGACAGGGCCAGCAGGAGACACC  -81
ATGGGGCACCTCTCAGCCCCACTTCACAGAGTGCGTGTACCCTGGCAGGGGCTTCTGCTCACAGCCTCACTTCTAACCTTCTGGAACCCG   90
M  G  H  L  S  A  P  L  H  R  V  R  V  P  W  Q  G  L  L  L  T  A  S  L  L  T  F  W  N  P
         -30                                -20                                -10
CCCACCACTGCCCAGCTCACTACTGAATCCATGCCATTCAATGTTGCAGAGGGGAAGGAGGTTCTTCTCCTTGTCCACAATCTGCCCCAG  180
P  T  T  A  Q  L  T  T  E  S  M  P  F  N  V  A  E  G  K  E  V  L  L  L  V  H  N  L  P  Q
      -1  1                          10                                 20
CAACTTTTTGGCTACAGCTGGTACAAAGGGGAAAGAGTGGATGGCAACCGTCAAATTGTAGGATATGCAATAGGAACTCAACAAGCTACC  270
Q  L  F  G  Y  S  W  Y  K  G  E  R  V  D  G  N  R  Q  I  V  G  Y  A  I  G  T  Q  Q  A  T
      30                                 40                                 50
CCAGGGCCCGCAAACAGCGGTCGAGAGACAATATACCCCAATGCATCCCTGCTGATCCAGAACGTCACCCAGAATGACACAGGATTCTAC  360
P  G  P  A  N  S  G  R  E  T  I  Y  P  N  A  S  L  L  I  Q  N  V  T  Q  N  D  T  G  F  Y
      60                                 70                                 80
ACCCTACAAGTCATAAAGTCAGATCTTGTGAATGAAGAAGCAACTGGACAGTTCCATGTATACCCGGAGCTGCCCAAGCCCTCCATCTCC  450
T  L  Q  V  I  K  S  D  L  V  N  E  E  A  T  G  Q  F  H  V  Y  P  E  L  P  K  P  S  I  S
      90                                100                                110
AGCAACAACTCCAACCCTGTGGAGGACAAGGATGCTGTGGCCTTCACCTGTGAACCTGAGACTCAGGACACAACCTACCTGTGGTGGATA  540
S  N  N  S  N  P  V  E  D  K  D  A  V  A  F  T  C  E  P  E  T  Q  D  T  T  Y  L  W  W  I
     120                                130                                140
AACAATCAGAGCCTCCCGGTCAGTCCCAGGCTGCAGCTGTCCAATGGCAACAGGACCCTCACTCTACTCAGTGTCACAAGGAATGACACA  630
N  N  Q  S  L  P  V  S  P  R  L  Q  L  S  N  G  N  R  T  L  T  L  L  S  V  T  R  N  D  T
     150                                160                                170
GGACCCTATGAGTGTGAAATACAGAACCCAGTGAGTGCGAACCGCAGTGACCCAGTCACCTTGAATGTCACCTATGGCCCGGACACCCCC  720
G  P  Y  E  C  E  I  Q  N  P  V  S  A  N  R  S  D  P  V  T  L  N  V  T  Y  G  P  D  T  P
     180                                190                                200
ACCATTTCCCCTTCAGACACCTATTACCGTCCAGGGGCAAACCTCAGCCTCTCCTGCTATGCAGCCTCTAACCCACCTGCACAGTACTCC  810
T  I  S  P  S  D  T  Y  Y  R  P  G  A  N  L  S  L  S  C  Y  A  A  S  N  P  P  A  Q  Y  S
     210                                220                                230
TGGCTTATCAATGGAACATTCCAGCAAAGCACACAAGAGCTCTTTATCCCTAACATCACTGTGAATAATAGTGGATCCTATACCTGCCAC  900
W  L  I  N  G  T  F  Q  Q  S  T  Q  E  L  F  I  P  N  I  T  V  N  N  S  G  S  Y  T  C  H
     240                                250                                260
GCCAATAACTCAGTCACTGGCTGCAACAGGACCACAGTCAAGACGATCATAGTCACTGAGCTAAGTCCAGTAGTAGCAAAGCCCCAAATC  990
A  N  N  S  V  T  G  C  N  R  T  T  V  K  T  I  I  V  T  E  L  S  P  V  V  A  K  P  Q  I
     270                                280                                290
AAAGCCAGCAAGACCACAGTCACAGGAGATAAGGACTCTGTGAACCTGACCTGCTCCACAAATGACACTGGAATCTCCATCCGTTGGTTC 1080
K  A  S  K  T  T  V  T  G  D  K  D  S  V  N  L  T  C  S  T  N  D  T  G  I  S  I  R  W  F
     300                                310                                320
TTCAAAAACCAGAGTCTCCCGTCCTCGGAGAGGATGAAGCTGTCCCAGGGCAACACCACCCTCAGCATAAACCCTGTCAAGAGGGAGGAT 1170
F  K  N  Q  S  L  P  S  S  E  R  M  K  L  S  Q  G  N  T  T  L  S  I  N  P  V  K  R  E  D
     330                                340                                350
GCTGGGACGTATTGGTGTGAGGTCTTCAACCCAATCAGTAAGAACCAAAGCGACCCCATCATGCTGAACGTAAACTATAATGCTCTACCA 1260
A  G  T  Y  W  C  E  V  F  N  P  I  S  K  N  Q  S  D  P  I  M  L  N  V  N  Y  N  A  L  P
     360                                370                                380
CAAGAAAATGGCCTCTCACCTGGGGCCATTGCTGGCATTGTGATTGGAGTAGTGGCCCTGGTTGCTCTGATAGCAGTAGCCCTGGCATGT 1350
Q  E  N  G  L  S  P  G  A  I  A  G  I  V  I  G  V  V  A  L  V  A  L  I  A  V  A  L  A  C
     390                                400                                410
TTTCTGCATTTCGGGAAGACCGGCAGGGCAAGCGACCAGCGTGATCTCACCAGCACAAACCCTCAGTCTCCAACCACACTCAGGACCACT 1440
F  L  H  F  G  K  T  G  R  A  S  D  Q  R  D  L  T  S  T  N  P  Q  S  P  T  T  L  R  T  T
     420                                430                                440
CCAATGACCCACCTAACAAGATGAATGAAGTTACTTATTCTACCCTGAACTTTGAAGCCCAGCAACCCACACAACCAACTTCAGCCTCCC 1530
P  M  T  H  L  T  R
     450
CATCCCTAACGCAGCAGTCTGCGCCCTTACACAAAATGACATCAAAAAAATCGTAGCCTTCTCCACTTCAAGTCAACTAGGACTCATAAT 1620
AGTTACAATCGGCATCAACCAACCACACCTAGCATTCCTGCACATCTGTACCCACGCCTTCTTCAAAGCCATACTATTTATGTGCTCCGG 1710
GTCCATCATCCACAACCTTAACAATGAACAAGATATTCGAAAAATAGGAGGACTACTCAAAACCATACCTCTCACTTCAACCTCCCTCAC 1800
CATTGGCAGCCTAGCATTAGCAGGAATACCTTTCCTCACAGGTTTCTACTCCAAAGACCACATCATCGAAACCGCAAACATTCTGCGGA 1889
```

FIG. 2. Nucleotide and deduced amino acid sequence of the BGP I cDNA. The open reading frame begins at nucleotide 1 and contains a leader peptide region (102 bp), an N-terminal domain (324 bp), an immunoglobulin-like domain (534 bp) consisting of two subdomains A and B, a BGP I-specific domain (324 bp) (domain A'), a membrane-spanning domain (72 bp), and a cytoplasmic domain (105 bp). The lengths of 5' and 3' untranslated regions are 79 and 428 bp, respectively. The boundaries of the domains are represented by the arrows. The amino acid sequences underlined show chymotryptic peptides that were analyzed. Cysteine residues and potential asparagine glycosylation sites are indicated by boxed letters and dots, respectively. The membrane-spanning domain is boxed.

clones were further screened with a 428-bp *Eco*RI/*Acc* I fragment from the N-terminal region of CEA obtained from a cDNA clone (unpublished data). Among 7 positives, clone 26 was found to code for an additional 67 bp past the 3′ end of clone 36. Using a 118-bp *Ava* II/*Eco*RI fragment from the 3′ end of clone 26 as a probe, we identified one clone (clone 4) representing the C-terminal coding region and parts of the 3′ untranslated region. Rescreening the same library with a 298-bp *Eco*RI/*Bgl* II fragment from clone 36 (Fig. 1) yielded 6 clones that were also positive with a 113-bp *Eco*RI/*Alu* I fragment from a CEA cDNA clone (unpublished data) encoding the leader peptide. Sequence analysis showed that clone 18 contained part of the 5′ untranslated region and the complete signal peptide of BGP I. The combined sequence of the overlapping cDNA clones is shown in Fig. 2.

A domain model for BGP I (see Fig. 1), based on previously determined structures for CEA and NCA (9, 16), predicts a 34-amino acid leader peptide, a 108-amino acid N-terminal domain, a 178-amino acid immunoglobulin-like domain, a 108-amino acid region specific to BGP I, a 24-amino acid membrane-spanning domain, and a 35-amino acid cytoplasmic portion. The potential transmembrane domain was determined by analyzing the hydrophobicity of the amino acid sequence according to the method of Kyte and Doolittle (17). Although the cytoplasmic region adjacent to the proposed membrane-spanning domain contains several basic amino acids, no membrane stop-transfer signals (18) were found. This domain model predicts that BGP I would have a short threonine-rich cytoplasmic domain, which has not been found in either CEA or NCA (9, 16). The 3′ untranslated region is unique to BGP I when compared to CEA, NCA, or other genes in release 53.0 of the GenBank data base. Since neither a typical polyadenylylation signal nor a poly(A) tail was found in our clones, the actual length of the 3′ untranslated region could not be assessed.

RNA Blot Hybridization Analysis. Previous work demonstrated that the use of coding region probes from CEA or NCA gives multiple bands on electrophoretic hybridizations due to their high degree of sequence homology (19). We have recently shown that probes derived from the 3′ untranslated regions of NCA and CEA cDNAs are specific for the respective genes (20). Therefore a 285-bp *Hinc*II/*Eco*RI fragment from the 3′ untranslated region of BGP I in clone 4 (Fig. 1) was used to evaluate the expression of the gene in different human tissues. The results of the analysis are shown in Fig. 3. Except for the leukemia cell line K562, which did not react with the probe, a predominant 2.8-kilobase (kb) message was found in all tissues and was accompanied by two weaker bands 4.1 and 2.1 kb in size (Fig. 3*B*). Among these, the 4.1-kb message was present in normal liver, placenta, the breast cancer cell line MCF 7, and very faintly in the liver metastasis of the colon tumor. The 2.1-kb band could be shown in all of these samples except for the liver metastasis tissue, where it was not detectable even after long exposure. We also hybridized the same set of RNAs to the coding region specific for BGP I (domain A′), using a 159-bp *Alu* I fragment from clone 4 (Fig. 1). After overnight exposure only the 4.1- and 2.1-kb bands could be seen in the normal liver, while exposure for a period of 3 days showed these bands very faintly also in the human placental RNA (Fig. 3*A*). Surprisingly, the 2.8-kb species was not detected in any of the tissues with the probe from domain A′ even after long exposure.

Southern Blot Analysis. The results of the Southern blot analysis using a probe for the 3′ untranslated region are provided in Fig. 4. A single band was observed for each of the

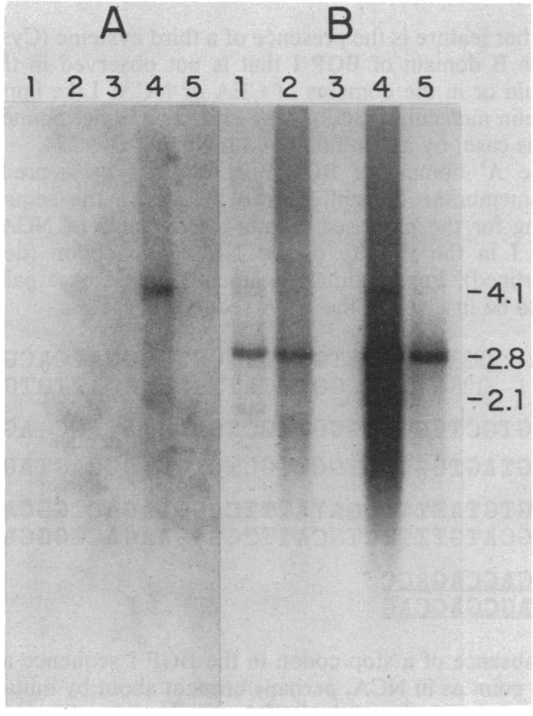

FIG. 3. Blot hybridization analysis of human RNAs with BGP I-specific probes. (*A*) The probe used was a ^{32}P-labeled BGP I unique coding region fragment (*Alu* I fragment as shown in Fig. 1). Lanes: 1, MCF7 mammary carcinoma; 2, human placenta; 3, K562 human leukemia; 4, human liver; and 5, liver metastasis of a human colon carcinoma. Total RNA (10 μg) was used for each lane except in the case of human liver, where poly(A)$^+$ RNA (3 μg) was used. (*B*) Lanes 1–5 as in *A*. The probe was a ^{32}P-labeled *Hinc*II/*Eco*RI fragment from the 3′-untranslated region of BGP I (see Fig. 1). After hybridization the final stringent wash was for 60 min in 0.1× SSC/0.1% NaDodSO$_4$ at 68°C. Numbers on the right are length in kb.

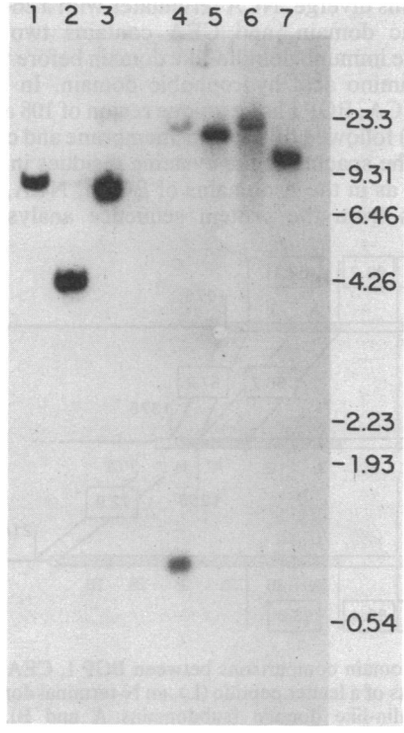

FIG. 4. Southern blot analysis of human DNA with a BGP I-specific probe from the 3′ untranslated region (*Hinc*II/*Eco*RI fragment as shown in Fig. 1). Human DNA was digested with the following endonucleases: lane 1, *Eco*RI; lane 2, *Eco*RI plus *Sst* I; lane 3, *Sst* I; lane 4, *Bam*HI; lane 5, *Pst* I; lane 6, *Pvu* II; and lane 7, *Hin*dIII. After hybridization the final stringent wash was for 60 min in 0.1× SSC/0.1% NaDodSO$_4$ at 68°C.

five restriction endonucleases used, with the exception of *Bam*HI. With this enzyme, a faint second band was seen at high molecular weight. This band is probably due to partial digestion of the DNA, since the probe does not contain an internal *Bam*HI site. These results are consistent with the presence of a single-copy gene as detected by this unique probe.

Protein Structural Studies. To confirm the amino acid sequence predicted by the cDNA clones, BGP I was purified from human bile and, after digestion with chymotrypsin, the sequences of eight peptides were determined (underlined in Fig. 2). Confirmation of the peptide sequence data was obtained by mass spectral analysis (data not shown). The peptide sequences, which covered 129 amino acids, were identical to the corresponding sequences predicted by the cDNA clones. The identity of the sequences confirmed that the cDNA clones coded for BGP I. The corresponding sequences in NCA and CEA contain 13 and 14 amino acid differences, respectively. In addition, all five of the predicted glycosylation sites in the peptides sequenced were confirmed. Repeated attempts to determine the N-terminal sequence of intact BGP I were unsuccessful, suggesting that the N terminus of BGP I is blocked.

DISCUSSION

A comparison of BGP I to NCA and CEA is presented in Fig. 5. The sequences of the leader peptide, the N-terminal domain, and the immunoglobulin-like domain (subdomains A and B) of BGP I are highly homologous to those of NCA and CEA. In addition, the location of the cysteine residues in the immunoglobulin-like domain, which are predicted to form intradomain disulfide bonds similar to those for the immunoglobulin gene superfamily members, are conserved. Following the immunoglobulin-like region, the structures of the three proteins diverge. NCA terminates with a 26-amino acid hydrophobic domain, and CEA contains two additional copies of the immunoglobulin-like domain before terminating with a 26-amino acid hydrophobic domain. In contrast to CEA and NCA, BGP I has a unique region of 108 amino acids (domain A') followed by the transmembrane and cytoplasmic domains. The spacing of the cysteine residues in domain A' is the same as in the A domains of BGP I, NCA, and CEA. The results from the protein sequence analysis (Fig. 2) confirm the sequences predicted by the cDNA data for regions in the N-terminal and immunoglobulin-like domain. The agreement between predicted and observed sequences, covering 129 residues, established that the isolated clones code for BGP I. In addition, the protein data identified the predicted glycosylation sites in all sequenced peptides. Repeated attempts to obtain an N-terminal sequence were unsuccessful, consistent with the conclusion that the N terminus of BGP I is blocked. The predicted cleavage site of the signal peptide from pro-BGP I would result in an N-terminal glutamine. Conversion of the glutamine to pyroglutamic acid would produce a blocked N terminus.

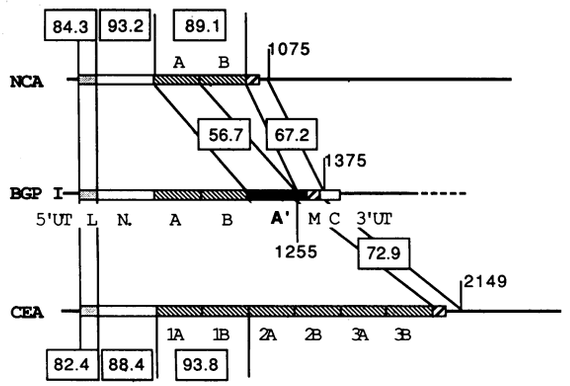

FIG. 5. Domain comparisons between BGP I, CEA, and NCA. BGP I consists of a leader peptide (L), an N-terminal domain (N), an immunoglobulin-like domain (subdomains A and B), a specific domain (A'), a membrane-spanning domain (M), and a cytoplasmic domain (C). The structures characteristic for BGP I are domains A' (closed box) and C (open box). Corresponding domains in BGP I, CEA, and NCA are shown in hatched, shaded, or open boxes. The boxed numbers represent percentage nucleotide identity between BGP I and the corresponding domains in NCA or CEA. The nucleotide sequences in BGP I, NCA, and CEA diverge at positions 1375, 1075, and 2149 in the respective genes as indicated.

The A' domain of BGP I is more similar to its A domain than to its B domain (57.4%). Surprisingly, the extent of identity between A' and A of BGP I is considerably less than that observed (80–90%) between the A domains of CEA (16). This difference may indicate that BGP I diverged from NCA with the acquisition of an ancestral immunoglobulin-like gene. It is thus postulated that, together with other members of the immunoglobulin gene superfamily, NCA evolved from a common ancestral immunoglobulin gene (9, 14, 19), after which BGP I and CEA diverged separately from NCA.

Further comparisons of the immunoglobulin-like domains of CEA, NCA, and BGP I with the A' domain show some distinct features. One of these occurs at the junction of the B domain and the A' domain. As shown below, there is a three-amino acid insertion in this junction compared to the 1B/2A junction of CEA:

```
BGP I   274–296   CNRTTVKTIIVTELSPVVAKPQI
CEA     274–293   LNRTTVTTITVY    AEPPKPFI
```

Another feature is the presence of a third cysteine (Cys-274) in the B domain of BGP I that is not observed in the A' domain or in the domains of CEA or NCA. Like immunoglobulin molecules, BGP I may exist as a dimer connected, in this case, by a disulfide bond involving Cys-274.

The A' domain of BGP I is followed by a predicted transmembrane domain. Shown below are the sequences coding for the proposed membrane domains of NCA and BGP I in the vicinity of the NCA stop codon (double-underlined). For maximal alignment a single base pair gap had to be inserted in the NCA sequence.

```
NCA   (983–1091)  TGGCCACCGTCGGCATCACGATC
BGP I (1284–1389) GGGCCATTGCTGGCATTGTGATC

GGAGTGCTGGCCAGGGTCGCTCTGATATAGC-AGCC

GGAGTAGTGGCCCTGGTTGCTCTGATAGCAGTAGCC

CTGGTGTATTTTCGATATTTCAGGAAGACTGGCAGA
CTGGCATGTTTTCTGCATTTCGGGAAGACCGGCAGG

TTGGACCAGACC
GCAAGCGACCAG
```

The absence of a stop codon in the BGP I sequence at the same point as in NCA, perhaps brought about by mutations during the evolution of the BGP I gene, explains the extended cytoplasmic domain of the BGP I protein sequence compared to NCA and CEA. Overall, the sequence identity observed in this region was 67% and 73%, respectively (Fig. 5). The identity between CEA and NCA for this same section is 86%. The lower similarity in this region for BGP I versus NCA compared to NCA versus CEA suggests that the BGP I gene arose from the NCA gene preceding the evolution of the CEA gene. Beyond this region, the sequences of BGP I, NCA, and CEA diverge in concert (underlined), suggesting that this may represent an exon/intron boundary. Considering the unique

properties of the NCA and CEA membrane domains, it is possible to speculate that this region constitutes a single exon. The absence of sequence homology between BGP I and NCA or CEA at the 5′ end of this section leaves this intron/exon boundary in doubt for BGP I.

The 3′ untranslated regions of the CEA and NCA genes have been shown to be specific for the respective genes (20). The results from the Southern blot analysis (Fig. 4) demonstrate that the 3′ untranslated region of BGP I is likewise specific and that the BGP I gene is present in a single copy. However, when the 3′ untranslated region was used to probe RNAs from different tissues, messages of three different sizes were observed (Fig. 3). A major band of 2.8 kb and weaker bands of 4.1 and 2.1 kb were observed in normal liver, placenta, a liver metastasis of a colonic carcinoma, and a breast carcinoma cell line (MCF7). The expression of the minor bands varied in the RNAs. The 2.1-kb message could not be seen and the 4.1-kb band was very faint in the metastasis tissue compared to the placenta and tumor cell line MCF7. Using the 159-bp *Alu* I fragment from the domain A′ in BGP I as a probe (Fig. 1), we showed that the 4.1- and the 2.1-kb messages hybridize, whereas the 2.8-kb band does not, indicating that the BGP I-specific domain A′ is not present in the 2.8-kb species. Previous studies performed on different forms of biliary glycoprotein and their relatedness to CEA and NCA involved polyclonal BGP I-specific antisera for isolation and characterization. These studies may explain the absence of the A′ domain from the 2.8-kb messenger RNA. For example, BGP I obtained from serum (s-BGP I) did not react with specific antisera against CEA and NCA, but it showed partial identity with BGP I isolated from bile, indicating that the two species differ in the number of BGP I-specific epitopes (3). This immunological difference could be explained with the elimination of an epitope-bearing domain by a posttranscriptional modification such as alternative RNA splicing. Cloning of the BGP I species represented by the 2.8-kb messenger RNA will help to assign the BGP I-specific antigenic determinants to the different domains in those molecules.

Our data therefore suggest that three different forms of BGP I can be generated from the single-copy gene by alternative splicing of a precursor RNA. There is plenty of evidence for posttranscriptional modification in the immunoglobulin gene superfamily. For example, the structure of BGP I is similar to that of neural cell adhesion molecule (N-CAM), another member of this family. In the case of N-CAM, three different forms of the C-terminal region have been described as a result of alternative RNA splicing (21). It is believed that these posttranscriptional events are involved in tissue-specific modulation of the function of N-CAM. While two forms of N-CAM are demonstrated to have a transmembrane and cytoplasmic domain, as is predicted for BGP I, the third species has a cell membrane insertion mechanism similar to that of CEA. Recently, it was shown that the mature form of CEA is processed posttranslationally to remove the proposed membrane domain with the subsequent addition of a glycosyl-phosphatidylinositol membrane anchor (22). Preliminary data suggests that NCA is anchored to the plasma membrane in a similar fashion (unpublished results). Comparison of the primary structures of N-CAM and BGP I leads to the hypothesis that the function(s) of BGP I on the cell surface of the bile canaliculi or in other tissues may be regulated in analogy to N-CAM, leading to a number of BGP I proteins translated from differentially processed RNA. The existence of a transmembrane and threonine-rich cytoplasmic domain could provide some insight into the possible function of BGP I. In contrast to CEA and NCA, the cytoplasmic domain of BGP I may allow it to interact with molecules in the cell cortex and to be further modulated by phosphorylation at the threonine residues. Considering the fact that we isolated BGP I from a human colon cDNA library and that the gene is transcribed in a number of human tissues, the question of tissue-specific expression of this CEA-like antigen remains to be solved. Clearly the liver is not the only organ that shows the active gene. Development of monoclonal antibodies with different epitope specificities to BGP I will help to clarify the tissue distribution and expression manner of BGP I during morphogenesis or in malignant transformation of the cell.

We are indebted to Dr. Steve Akman (Dept. of Medical Oncology, City of Hope) for the generous gift of the tumor cell lines K562 and MCF7. We gratefully acknowledge the support of National Institutes of Health Grant CA 37808 and Deutsche Forschungsgemeinschaft Grants WA-473/4-1 and NE-331/1-1.

1. Svenberg, T. (1976) *Int. J. Cancer* **17**, 588–596.
2. Neumaier, M., Fenger, U. & Wagener, C. (1985) *J. Immunol.* **135**, 3604–3609.
3. Svenberg, T., Hammerström, S. & Hedin, A. (1979) *Mol. Immunol.* **16**, 245–252.
4. Svenberg, T., Wahren, B. & Hammerström, S. (1979) *Clin. Exp. Immunol.* **36**, 317–325.
5. Davis, L. G., Dibner, M. D. & Battey, J. F. (1986) in *Basic Methods in Molecular Biology* (Elsevier, Amsterdam), pp. 185–189.
6. Feinberg, A. P. & Vogelstein, B. (1983) *Anal. Biochem.* **132**, 6–13.
7. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY).
8. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
9. Neumaier, M., Zimmermann, W., Shively, L., Hinoda, Y., Riggs, A. D. & Shively, J. E. (1988) *J. Biol. Chem.* **263**, 3202–3207.
10. Meese, E. & Blin, H. (1987) *Gene Anal. Tech.* **4**, 45–49.
11. Thomas, P. S. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5201–5205.
12. Kan, Y. W., Dozy, A. M., Trecartin, R. & Todd, D. (1977) *N. Engl. J. Med.* **297**, 1081–1084.
13. Wahl, G. M., Stern, M. & Stark, G. R. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3683–3687.
14. Paxton, R. J., Mooser, G., Pande, H., Lee, T. D. & Shively, J. E. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 920–924.
15. Zimmermann, W., Ortlieb, B., Friedrich, R. & von Kleist, S. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 2960–2964.
16. Oikawa, S., Nakazato, H. & Kosaki, G. (1987) *Biochem. Biophys. Res. Commun.* **142**, 511–518.
17. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105–132.
18. Sabatini, D., Kreibich, G., Morimoto, T. & Adesnik, M. (1982) *J. Cell Biol.* **92**, 1–22.
19. Thompson, J. A., Pande, H., Paxton, R. J., Shively, L., Padma, A., Simmer, R. L., Todd, C. W., Riggs, A. D. & Shively, J. E. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 2965–2969.
20. Zimmermann, W., Weber, B., Ortlieb, B., Rudert, F., Schempp, W., Fiebig, H., Shively, J. E., von Kleist, S. & Thompson, J. A. (1988) *Cancer Res.* **48**, 2550–2554.
21. Cunningham, B. A., Hemperly, J. J., Murray, B. A., Prediger, E. A., Brackenburg, R. & Edelman, G. M. (1987) *Science* **236**, 799–806.
22. Hefta, S. A., Hefta, L. J. F., Lee, T. D., Paxton, R. J. & Shively, J. E. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4648–4652.