Molecular Cloning and Expression of Human Tumor-associated Polymorphic Epithelial Mucin*

Sandra J. Gendler†, Carole A. Lancaster, Joyce Taylor-Papadimitriou, Trevor Duhig, Nigel Peat, Joy Burchell, Lucy Pemberton, El-Nasir Lalani, and David Wilson

From the Imperial Cancer Research Fund, Lincoln's Inn Fields, London WC2A 3PX, United Kingdom

Human mammary cells present on the cell surface a polymorphic epithelial mucin (PEM) which is developmentally regulated and aberrantly expressed in tumors. PEM carries tumor-associated epitopes recognized by the monoclonal antibodies HMFG-1, HMFG-2, and SM-3. Previously isolated partial cDNA clones revealed that the core protein contained a large domain consisting of variable numbers of 20-amino acid repeat units. We now report the full sequence for PEM, as deduced from cDNA sequences. The encoded protein consists of three distinct regions: the amino terminus consisting of a putative signal peptide and degenerate tandem repeats; the major portion of the protein which is the tandem repeat region; the carboxyl terminus consisting of degenerate tandem repeats and a unique sequence containing a transmembrane sequence and a cytoplasmic tail. Potential O-glycosylation sites (serines or threonines) make up more than one-fourth of the amino acids. Length variations in the tandem repeat result in PEM being an expressed variable number tandem repeat locus. Tandem repeats appear to be a general characteristic of mucin core proteins.

Mucins are large molecular weight glycoproteins which contain at least 50% carbohydrate O-linked through N-acetylgalactosamine to serine and/or threonine. Recently, attention has been focused on mucin glycoproteins because many antibodies, selected for the specificity of their reactions with normal and/or malignant epithelial cells, recognize epitopes on these complex molecules. However, although a certain amount of data is available on the structure of the carbohydrate side chains of some mucins, little is known about the primary structure of the core proteins. We and others (Gendler et al., 1987a; Siddiqui et al., 1988) have recently isolated cDNA clones coding for a domain of the core protein of a polymorphic epithelial mucin (PEM) which is expressed by breast and other carcinomas and is found in human milk. The cDNA clones, which were isolated by screening an expression library with antibodies directed to the core protein, were found to consist of varying numbers of a conserved tandem repeat coding for a 20-amino acid unit (Gendler et al., 1988). Of particular interest was the extensive polymorphism attributable to different numbers of the tandem repeat (Swallow et al., 1987). The amino acid sequence of the tandem repeat is rich in serines and threonines and therefore is likely to be highly glycosylated. A tandemly repeated sequence has also been found to form the basis for a structural domain of other mucins. Partial cDNA clones coding for the core proteins of the porcine submaxillary mucin (Timpte et al., 1988) and the human colonic mucin (Gum et al., 1989) consist of tandem repeats of 243 and 69 base pairs (bp), respectively. Although the sequences differ from the sequence of the PEM tandem repeat, they too contain high levels of serine and/or threonines and are presumably heavily glycosylated.

The polymorphic epithelial mucin is of particular interest, since it is expressed by breast and other carcinomas in an aberrantly glycosylated form (Burchell et al., 1987). It appears that the carbohydrate side chains of the cancer-associated mucin are shorter than the side chains of the mucin produced by normal cells (Hull et al., 1988; Hanisch et al., 1989). This may result in the exposure of peptide epitopes on the cancer cell mucin which are masked in the fully glycosylated form (Burchell et al., 1989). Other epitopes in the tandem repeat domain are expressed in both the normally processed and cancer-associated mucin (for review, see Burchell and Taylor-Papadimitriou, 1989).

Since PEM is a tumor-associated mucin (Girling et al., 1989) and is found in the serum of cancer patients (Burchell et al., 1984), many of the antibodies reactive with the mucin are in use as diagnostic agents and even in therapeutic studies (Hilkens et al., 1986; Epenetos et al., 1985; Granowska et al., 1984; Hammersmith Oncology Group and Imperial Cancer Research Fund, 1984). Thus, there is great interest in defining the full structure of the core protein of PEM as an example of a mucin glycoprotein which is of clinical importance. Here we report the sequence of a full-length cDNA coding for the core protein of the polymorphic epithelial mucin. The presence of a putative signal sequence at the 5′ end of the message and a putative transmembrane sequence at the 3′ end suggest that the message codes for a membrane-anchored form of the mucin. That this is indeed the case was shown by demonstrating membrane staining in COS cells transfected with an expression vector containing the full-length cDNA. Since one form of the mucin exists as a transmembrane protein, it may be an important target antigen for those types of immunotherapy which require internalization of the antibody. Moreover, the presence of a cytoplasmic tail of 69 amino acids suggests that the mucin may play some role in signal transduction or cellular organization.

MATERIALS AND METHODS

Synthesis of 5′ cDNA Using Anchored Polymerase Chain Reaction—The anchored polymerase chain reaction procedure (Loh et al., 1990)
Thirteen clones were picked and purified, the largest of which probes were 3' end-labeled to a specific activity of 3 x 10^6 dpm/µg gene; the nucleotide sequence was determined by the dideoxychain termination method (Sanger et al., 1980). The library of 2 x 10^6 colonies was screened using Church and Gilbert random hexamers were used as primers and plated onto DH10b cells and washed according to the manufacturer's instructions, (pGEM-PEM17) contained about 200 bp of tandem repeat sequence of the cosmid library, kindly contributed by Dr. A.-M. Frischauf (Imperial Cancer Research Fund, London), was digested with PvuII site together with the tandem repeat (GCTGGTGG) of X phage which has been implicated as a hot spot for RecA-mediated recombination in Escherichia coli. Since a part of the tandem repeat sequence of the PEM core protein was determined by the dideoxy chain termination method, the nucleotide sequence at the 5' terminus and clones with 3' sequences including the unique 3' sequence previously reported by us; sequence at the 5' terminus and clones with 3' sequences were obtained by digesting the genomic cosmid clone GPEM-1 with KpnI to liberate a 2300-bp fragment containing approximately 2100 bp. The 3' portion of the construct was obtained by digesting the 3'-clone pGEM-PEM21 with Clal (a 3'-polylinker) with fluorescein isothiocyanate-conjugated rabbit antionmune immunoglobulins (Dako Corp., Santa Barbara, CA).

RESULTS

Cloning Strategy—In an attempt to obtain a full-length cDNA clone or clones with sequence 5' to the tandem repeat, two Agt10 libraries were screened with a probe for the tandem repeat pMUC10 (Gendler et al., 1987a). The cosmid contributed by Dr. D. W. Russell (University of Texas at Dallas). These vectors contain the SV40 origin, the cytomegalovirus major immediate early gene promoter (and in pCMV4, the early region enhancer), and the human growth hormone fragment with transcription termination and polyadenylation signals. The construct was verified as before. These plasmids will be referred to as pCMV4-P pem-tm and pCMV5-P pem-tm.

DNA Electroporation and Expression—Plasmid DNA was introduced into COS cells by electroporation using a Bio-Rad gene pulser transfection apparatus at 450 V with a capacitance of 250 microFarads using 35 µg of plasmid DNA/10^6 cells in phosphate-buffered saline, pH 7.0. Cells were assayed for PEM expression at 65 h after electroporation by staining with the monoclonal antibodies HMFG-1, HMFG-2, and SM-3 which react with epitopes within the tandem repeat of the PEM core protein (Gendler et al., 1988). Binding of the antibodies was visualized either with peroxidase-conjugated rabbit antimouse immunoglobulins or with fluorescein isothiocyanate-conjugated rabbit antimouse immunoglobulins.

acid sequences were performed on a VAX computer using Intelligenetics software (Mountain View, CA).

Construction of full-length cDNA Used for Expression—Fragments from three separate plasmids were used to construct the hybrid plasmid used in our expression studies. The plasmid clone pBS 5'PEM, containing a 458-bp insert from the 5' cDNA made using polymericase chain reaction was digested with BamHI and PvuII to yield a 1053-bp fragment which was cut out electrophoretically from the BamHI restriction site to the PvuII site at base 345. The tandem repeat portion was obtained by digesting the genomic cosmid clone GPEM-1 with KpnI to liberate a 2300-bp fragment containing approximately 2100 bp. The rest of the fragment was purified from an agarose gel and, in order to promote the correct orientation during ligation, the 5' and 3' cDNA fragments were phosphorylated and treated with alkaline phosphatase. The three fragments were ligated into a pUC vector and transformed into DH10b cells. The correct orientation of DNA was ascertained by restriction digests and by sequencing all the ligation junctions.

The full-length cDNA of 3762 bp was cloned into pCMV4 and pCMV5 mammalian expression vectors (Andersson et al., 1989) kindly supplied by Dr. D. W. Russell (University of Texas at Dallas). These vectors contain the SV40 origin, the cytomegalovirus major immediate early gene promoter (and in pCMV4, the early region enhancer), and the human growth hormone fragment with transcription termination and polyadenylation signals. The construct was verified as before. These plasmids will be referred to as pCMV4-P pem-tm and pCMV5-P pem-tm.

DNA Electroporation and Expression—Plasmid DNA was introduced into COS cells by electroporation using a Bio-Rad gene pulser transfection apparatus at 450 V with a capacitance of 250 microFarads using 35 µg of plasmid DNA/10^6 cells in phosphate-buffered saline, pH 7.0. Cells were assayed for PEM expression at 65 h after electroporation by staining with the monoclonal antibodies HMFG-1, HMFG-2, and SM-3 which react with epitopes within the tandem repeat of the PEM core protein (Gendler et al., 1988). Binding of the antibodies was visualized either with peroxidase-conjugated rabbit antimouse immunoglobulins or with fluorescein isothiocyanate-conjugated rabbit antimouse immunoglobulins.

acid sequences were performed on a VAX computer using Intelligenetics software (Mountain View, CA).

RESULTS

Cloning Strategy—In an attempt to obtain a full-length cDNA clone or clones with sequence 5' to the tandem repeat, two Agt10 libraries were screened with a probe for the tandem repeat pMUC10 (Gendler et al., 1987a). The cosmid contributed by Dr. D. W. Russell (University of Texas at Dallas). These vectors contain the SV40 origin, the cytomegalovirus major immediate early gene promoter (and in pCMV4, the early region enhancer), and the human growth hormone fragment with transcription termination and polyadenylation signals. The construct was verified as before. These plasmids will be referred to as pCMV4-P pem-tm and pCMV5-P pem-tm.

DNA Electroporation and Expression—Plasmid DNA was introduced into COS cells by electroporation using a Bio-Rad gene pulser transfection apparatus at 450 V with a capacitance of 250 microFarads using 35 µg of plasmid DNA/10^6 cells in phosphate-buffered saline, pH 7.0. Cells were assayed for PEM expression at 65 h after electroporation by staining with the monoclonal antibodies HMFG-1, HMFG-2, and SM-3 which react with epitopes within the tandem repeat of the PEM core protein (Gendler et al., 1988). Binding of the antibodies was visualized either with peroxidase-conjugated rabbit antimouse immunoglobulins or with fluorescein isothiocyanate-conjugated rabbit antimouse immunoglobulins. The full-length cDNA clone or clones with sequence 5' to the tandem repeat, two Agt10 libraries were screened with a probe for the tandem repeat. All the clones obtained lacked any nonrepetitive sequence at the 5' terminus and clones with 3' sequences (including the unique 3' sequence previously reported by us; Gendler et al., 1988) appeared to be rearranged. Thus, a different strategy was adopted. To obtain the 5' sequence we synthesized cDNA corresponding to the 5' end of a breast cancer cell line transcript using an anchored polymerase chain reaction. The 3' cDNA clones were obtained by screening a plasmid library with oligonucleotide probes made to the unique 3' sequence obtained from genome clones. A plasmid library, grown in DH1α cells (RecA−), was used instead of a λ library, because of the possibility of recombination occurring when λ is grown in RecA + cells. This recombination might have been expected since a part of the tandem repeat sequence (GCTGGTGG) is closely related to the chi sequence (GCTGGTTG) of λ phage which has been implicated as a hot spot for RecA-mediated recombination in Escherichia coli.

Anchored Polymerase Chain Reaction Synthesis of 5' cDNA—The 5' terminus of the PEM core protein was synthesized using anchored polymerase chain reaction. The first strand of cDNA was synthesized using a breast cancer cell
line (BT20) transcript and a primer to the tandem repeat (see "Materials and Methods"). Amplification was carried out and a single band of DNA of about 550 bp was obtained and purified from an agarose gel, digested with restriction enzymes, and ligated into pBS-SK+. Four colonies were selected for sequencing, and the sequences agreed with each other and with sequences obtained from genomic clones. A leader sequence of 72 bp preceded the first ATG which was in-frame with the reading frame of the tandem repeat as previously determined (Fig. 1), and the sequence preceding this first ATG, CCACATGTA, is in close agreement with the Kozak (1989) consensus sequence except for the +4 position which is an A instead of a G.

The primer extension technique was used to precisely map the position of the cap site. An oligonucleotide extending from base 93 to 73 (ending at the A of the ATG) was labeled with [α-32P]dATP and elongated in the presence of H120 RNA with reverse transcriptase and unlabelled dNTPs to yield two bands 72 and 71 bases upstream of the ATG (Fig. 2). The most prominent product was 72 bp, equal to the number of base pairs from the 5' end of the oligonucleotide primer to the 5' end of the polymerase chain reaction-derived clone, thus confirming that the cDNA represents the entire length of its corresponding cellular mRNA 5' to the tandem repeat. The presence of a second band may be due to interference with reverse transcriptase by secondary structure. Under identical conditions, no primer extension product was seen using RNA from Daudi cells which do not express the PEM mucin. Preceding the transcription start site by 24 bp is a TATAA box and multiple G/C boxes.

Isolation of 3' cDNA Clones—A human genomic library in pCOS2EMBL was screened with a probe to the tandem repeat. One clone designated GPEM1 was selected and characterized by restriction mapping. Sequence analysis of the 389-bp KpnI-BarnHI band from the 3' end of the gene revealed an open reading frame. This band hybridized to RNA transcripts of equal size to those which hybridized with the tandem repeat sequence in breast cancer cell lines expressing the PEM gene (Fig. 3A). Overlapping oligonucleotides (7mers) were synthesized, 3' end-labeled to high specific activity, and used to screen a cDNA library constructed from BT20 poly(A) RNA in the plasmid pGEM-7Zf+. Four colonies were selected, the largest of which (pGEM-PEM17) contained about 300 bp of tandem repeat sequence as well as a poly(A) tail within its 1600 bp. pGEM-PEM17 was fully sequenced and the remaining clones were sequenced sufficiently to determine that no obvious rearrangements had occurred and that the 3'-nonrepetitive sequences were the same in the three largest clones (Fig. 3B).

To confirm that the polymerase chain reaction-derived clone and the 3' cDNA clones were part of the same gene, Northern blots were probed with cDNAs corresponding to the 5', tandem repeat, and 3' portions of the PEM cDNA (Fig. 3A). An identical pattern of hybridization was observed using the three probes, thus indicating their origin from the same transcript.

**Nucleotide Sequence of cDNA Clones—**Fig. 1 shows the composite DNA sequence from the 5' anchored polymerase chain reaction-derived clone, the consensus sequence of the tandem repeat, and the 3' cDNA clone. Sequences were determined in both directions for both the 5' and 3' portions of the cDNA. The region of conserved tandem repeats was not sequenced in full, although a cDNA tandem repeat clone obtained previously had been circularized, sonicated, and about 40 clones sequenced (Gendler et al., 1988). The vast majority of the 60-bp repeat units contained the consensus

---

2 C. Lancaster, manuscript in preparation.
DNA levels (Gendler et al., 1987a; Swallow et al., 1987). The VNTR alleles are detected by a number of enzymes that cut DNA outside of the tandem repeat, including BamHI, EcoRI, KpnI, PstI, Aul, HinfI, RsaI, Mbol, NeoI, and PvuII (Gendler et al., 1989). To determine the size variations of the tandem repeat region that exist, genomic DNAs from 69 random individuals, mainly of Northern European extraction, were digested with the restriction enzyme HinfI which cuts 932 bp from the 5' beginning of the tandem repeat and 586 bp from the 3' end, resulting in relatively small alleles. No individual exhibited more than two alleles. Based on their relative migration the observed fragments range in size from about 3 to about 9 kb (data not shown). Fig. 4 shows the distribution of the 30 different alleles detected among the 69 individuals. The most frequent allele is the 4000-bp allele that contains 41 repeat units. The next most frequent allele was the 6600-bp allele with 85 repeat units (Fig. 4). Not surprisingly, the most common genotype observed (five out of 69) is the heterozygote consisting of the two most common alleles.

Predicted Amino Acid Sequence and Composition of the PEM Core Protein—The core protein amino acid composition is dominated by the amino acid composition of the tandem repeat. Serine, threonine, proline, alanine, and glycine account for about 60% of the amino acids. The amino acid composition of the various domains is typical of mucin glycoproteins (Fig. 5).

The deduced sequence of the PEM core protein, which has the features of an integral membrane protein, consists of three distinct regions (see Figs. 1 and 5): (a) the amino-terminal region containing a hydrophobic signal sequence and degenerate tandem repeats; (b) the tandem repeat region

---

**Fig. 2. Mapping of the PEM mRNA 5' end.** A synthetic oligonucleotide (21mer) complementary to nucleotides +73 to +93 was end-labeled with [γ-32P]ATP and T4 polynucleotide kinase and incubated with 40 μg of total RNA from BT20 (lane 1) and Daudi (lane 2) cells. After primer extension by reverse transcriptase, products were separated on a 6% sequencing gel. The fragments (GCAT) of a sequencing reaction served as size markers. The primary extension product is 72 bases from the ATG translation start site.

**Fig. 3. Human PEM mRNA and restriction enzyme map of cDNA.** A, RNA transfer blot hybridization. Ten μg of total RNA from T47D (lane 1), BT20 (lane 2) and Daudi (lane 3) cells were denatured by glyoxal and urea, electrophoresed in a 1.4% agarose gel, transferred to Biodyne nylon membranes, and probed with 5', tandem repeat, and 3' cDNA probes (designated a, b, and c, in B). The two bands detected in lane 1 measure 6400 and 4700 bp and represent products from two codominant alleles. BT20 breast cancer cells which are homozygous for PEM (Gendler et al., 1987b) exhibit a single band in lane 2 of about 4800 bp. Daudi cells do not express PEM. Exposure time for probe b was a quarter of that for probes a and c, to enable distinct bands to be seen. The tandem repeat probe gives a much stronger signal on blots than do probes to unique sequences. B, human PEM, as deduced from cDNA and genomic clones, is depicted with a single tandem repeat which is defined by the first two Smal sites. All of the sites for the restriction enzymes, Pv (PvuII), S (Smal), St (StyI), K (KpnI), P (PstI), and B (BamHI) are shown. Below this line, the horizontal bars indicate the extent of the characterized cDNA clones. Both strands of A-PCR and pGEM-PEM17 were sequenced to completion. Hatched bars on the bottom line depict the region of the genomic clone that was sequenced to obtain a unique probe for the 3' cDNA.
and (c) the carboxyl-terminal region containing degenerate tandem repeats, a hydrophobic putative membrane spanning cytoplasmic tail.

At the amino terminus a putative signal peptide of 13 amino acids follows the first 7 amino acids. However, the actual site of cleavage has not been determined as attempts to obtain the amino-terminal sequence of the core protein were hindered by a blocked amino terminus. Following the signal sequence and preceding the first SMaI site (which we have used to define the beginning of the tandem repeat region) are 107 amino acids. Greater than 50% of these amino acids comprise degenerate tandem repeats (Fig. 1).

Since the number of tandem repeats per molecule is large (greater than 21 for the smallest allele we have observed), this domain forms the major part of the core protein, and results in a highly repetitive structure which is extremely immunogenic (Gendler et al., 1988). The sequence of the 20-amino acid tandem repeat unit corresponds to what might be expected for a protein which is extensively N-glycosylated. Five potential glycosylation sites are found in the repeat and these potential glycosylation sites are separated by regions rich in prolines (see Fig. 1).

As with the 5' end, the immediate sequence 3' to the last SMaI site of the conserved tandem repeat is made up of degenerate tandem repeats which lead into novel unique sequences. Residues 375-405 form a hydrophobic sequence which is a putative transmembrane domain. This being so, there appears to be a sizable cytoplasmic tail of 69 amino acids, beginning with a cluster of basic amino acids (Arg-Arg-Lys) immediately following the putative transmembrane sequence. The locations of five possible N-glycosylation sites (Asn-X-Ser/Thr) in the extracellular domain 3' to the tandem repeat region are indicated in Fig. 1.

**Construction of PEM cDNA**—Since full-length cDNA clones were not obtained from either the Agt10 library of the plasmid library, a full-length clone was constructed. The hybrid cDNA is composed of: (a) 346 bp of 5' PEM cDNA sequence extending from the transcription start site to the PvuII site prior to the tandem repeat; (b) the tandem repeat coding sequence from the PvuII site to the first Styl site 3' of the tandem repeat, obtained from the GPEM1 cosmid clone of human genomic DNA (part of exon 2); and (c) the 1306-bp fragment extending from the Styl site to the ClaI site in the pGEM polylinker which makes up the 3' coding and noncoding sequence and polyadenylation signal in clone pGEM-PEM17. Using both total and partial restriction enzyme digests to obtain the appropriate portions of the gene (detailed under "Materials and Methods"), a full-length cDNA was constructed and cloned into pBS-KSII+.

**Expression of PEM cDNA**—The constructed cDNA was cloned into the mammalian expression vectors pCMV-4 and pCMV-5 (see "Materials and Methods"). To express the PEM cDNA, the pCMV-PEM-tm clone was transfected by electroporation into COS cells. Indirect immunoperoxidase and immunofluorescence staining of fixed and unfixed cells using the monoclonal antibodies HMFG-1, HMFG-2, and SM-3 which are specific for the 20-amino acid tandem repeat revealed the presence of the PEM mucin in ~25% of the electroporated cells. All three antibodies gave a similar pattern of staining which is illustrated in Fig. 6. COS cells that were transfected with pCMV-4 or pCMV-5 containing no insert showed no staining with any of the antibodies (data not shown). To verify that the mucin was expressed in the membrane, immunofluorescence staining of unfixed electro-
**Core Protein Structure of a Carcinoma-associated Mucin**

amino acid composition of PEM is typical of that for a mucin, with serine, threonine, proline, glycine, and alanine accounting for greater than 60% of the amino acids. Previous studies have presented partial cDNA sequences comprised of a precise 60-bp tandem repeat (Gendler et al., 1988; Siddiqui et al., 1988). Analysis of the full-length cDNA revealed that the mucin core protein consists largely of tandem repeats which would allow for up to one-fourth of the amino acids (the serines and threonines) to be glycosylated (Fig. 5). Common sizes of the tandem repeat portion of the molecule occurring in an unrelated population of individuals are 820 and 1700 amino acids; the remainder of the protein consists of 480 amino acids, some of which are actually degenerate repeats occurring at the ends of the repeat domain. The 3' portion of the cDNA contains a region coding for a putative 31-amino acid transmembrane segment which would result in 100 amino acids within the membrane and cytoplasm. The peptide backbone of about 120,000–225,000 Da is in good accord with a protein of apparent of 240,000–450,000-Da mass containing about 50% by weight carbohydrate.

It is noteworthy that the PEM mucin sequence is not statistically homologous with any other protein or DNA sequence registered in the EMBL, GenBank, or Pir data bases. However, cDNA clones with sequences corresponding to the tandem repeat domain of the PEM gene have been isolated by three other groups. In two cases, clones were isolated from AΓ11 expression libraries constructed from MCF-7 cells or from T47D cells using antibodies DF3 (Siddiqui et al., 1988) or H23, respectively (Wreschner et al., 1989). Both of these antibodies were developed against breast cancer cells or their products. Surprisingly, however, clones selected by antibodies developed to the core protein of a pancreatic mucin (secreted by the HPAF cell line) also contained the tandem repeat region of the PEM gene (Lan et al., 1990), suggesting a similarity between the core proteins of the breast-associated mucin and the pancreatic mucin. In fact, the core proteins appear to be identical since the 4-kb cDNA clone selected from the HPAF library differs from the PEM gene sequence reported here only at bp 793 (T for an A, Ser for Thr), bp 1440 (C for T, no amino acid change), and bp 1483 (G for A, Ala for Thr) and 4 nucleotide changes in the 3'-untranslated region.

Although the full-length cDNA coding for PEM has not previously been reported, partial sequences 5' and 3' of the tandem repeat domain have been reported for the DF-3 antigen (Abe et al., 1989; Merlo et al., 1989). These sequences are similar to the corresponding regions of the PEM cDNA, but there are differences. The 5' sequence differs from the sequence reported here and from that determined by Lan et al. (1990) in having 27 extra base pairs (coding for the amino acid sequence Ala-Thr-Thr-Ala-Pro-Lys-Pro-Ala-Thr inserted between bp 129 and 130. Since this sequence is found in the first intron of our genomic clones, it would appear that the mRNA identified by Abe and colleagues (1989) was produced by alternative splicing, using a different AG as splice acceptor.

In the 3' sequence of Merlo et al. (1989), the DF-3 sequence is similar to our Fig. 1 sequence. However, there are numerous sequencing differences (e.g. GC for CG, CC for CCC) and a 14-bp duplication (bases 772–785 are shown twice in succession in the DF-3 sequence) which result in an early stop codon at base 768 (according to our sequence) and no similarity in translated amino acids in the 3' region following the tandem repeats.

Allelic variations in length have been described for several proteins with tandem repeat domains but, with the exception

---

**Fig. 6. Expression of PEM cDNA in COS cells.** Full-length PEM cDNA was cloned into the mammalian expression vector pCMV-5 (see "Materials and Methods") and the plasmid transfected into COS cells by electroporation. Cells were analyzed for PEM expression by staining with the antibody HMFG-2 and the binding visualized using a fluorescein isothiocyanate-conjugated rabbit antibody. A, HMFG-2 on methanolacetone-fixed cells and B, corresponding phase contrast. C, HMFG-2 on live cells and D, corresponding contrast.

**DISCUSSION**

We report here the full-length sequence of tumor-associated polymorphic epithelial mucin, PEM. Our data show that the
of the proline-rich proteins (Azen et al., 1984) and other mucins (Timpte et al., 1988; Gum et al., 1989), they are found in lower organisms (Muscavitch et al., 1982; Manning and Gaze, 1980; Ozaki et al., 1983; Sorimachi et al., 1988). The tandem repeat domain in the PEM gene shows allelic variations in length which result in such a high degree of polymorphism that the sequence can be considered a variable number tandem repeat (VNTR) locus (Swallow et al., 1987). A number of other VNTR loci have been identified by other investigators (Jeffreys et al., 1985; Nakamura et al., 1987); however, these loci are not expressed. Because most individuals will be heterozygous, VNTR loci are valuable genetic markers for human linkage maps, loss of heterozygosity studies, forensic analysis, and parentage testing, 69 unrelated individuals were examined and 30 different alleles were detected. The observed heterozygosity was 80%. This number is likely to be a low estimate, as our gels did not resolve DNA differing by <100 bp. This heterozygosity makes PEM a useful locus to study, particularly in light of its location on chromosome 1q21, a region that is frequently found to be altered in cancer. Indeed, analysis of paired samples of DNA prepared from breast cancers and blood cells from the same patient has shown that 30% of breast cancers show loss of an allele and 1% exhibited a new third band (Gendler et al., 1990). Not surprisingly, similar allele losses have been observed using probes for the gene coding for the DF3 antigen (Merlo et al., 1989) since this is the same gene as that coding for PEM. It should be noted that in the original experiments showing that the allelic variations in PEM were inherited in an autosomal codominant fashion, were done using lectins and antibodies to detect alleles of a mucin found in urine. When the locus was mapped to chromosome 1q21, it was given the designation TUM (peanut urinary mucin). Clearly, the mucin found in urine contains the same core protein as the breast and pancreatic mucin, since it reacts with PEM core protein-reactive antibodies and shows variations in size which accord with the variations in the restriction fragment length polymorphism shown by the PEM gene.

It is interesting to note that in other large structures with a high content of O-linked carbohydrate, exact repeats of short stretches of amino acids also occur. This is the case for the human intestinal mucin (Gum et al., 1989), the porcine submaxillary gland mucin (Timpte et al., 1988), and the polysialoglycoproteins of Rainbow trout eggs (Sorimachi et al., 1988). The variations in sizes of the molecules suggest that the length is not crucial to the function, but rather that the core exists in an extended form as a scaffold for the O-linked carbohydrate. Although no sequence motif has been identified as the acceptor in O-linked glycosylation, it has been suggested that prolines must reside near the serines and threonines (Briend et al., 1981; Hanover et al., 1980) and that these two amino acids should be adjacent to themselves or one another to be glycosylated (Timpte et al., 1988; Aubert et al., 1976). The PEM tandem repeat consists of 25% proline and four out of the five serine and threonine residues in the tandem repeat are located adjacent in the sequence. The tandem repeat structure, therefore, has the potential to be highly glycosylated. A similar pattern of distribution of prolines and adjacent serines and threonines is found in the tandem repeat sequences of the human intestinal mucin (Gum et al., 1989), whereas the porcine submaxillary gland mucin, although exhibiting a predominance of adjacent serine and threonine residues, contains only 6% proline (Timpte et al., 1988). While the sites of glycosylation are determined, at least in part by the amino acid sequence, the detailed structure of the carbohydrate side chains apparently shows tissue-specificity, since these side chains are very different in the PEM produced by breast and by pancreatic tumor cell lines. This means that the glycosylated mucins from the two tissues show dramatically different profiles of epitopes in the carbohydrate side chains which are more complex in the pancreatic mucin and mask core protein epitopes which are exposed in PEM produced by both normal and malignant breast epithelium.

In addition to the many potential O-glycosylation sites, the deduced amino acid sequence described here contains five potential N-glycosylation sites. This was not unexpected since the human mammary mucin has recently been shown to undergo N-glycosylation (Hilkens and Buijs, 1988; Linsley et al., 1988), although previously published oligosaccharide analysis (Schinizu and Yausuchi, 1982) shows no detectable mannose. Other mucins contain detectable quantities of N-linked carbohydrate, including mouse submandibular mucin (Denny and Denny, 1982; Amerongen et al., 1983), and potential N-glycosylation sites have been found in the human intestinal mucin (Gum et al., 1989) and the porcine submaxillary gland mucin (Timpte et al., 1988).

Some light may be thrown on the question of glycosylation sites from studies mapping core protein epitopes recognized by monoclonal antibodies. Some core protein epitopes are exposed in both the mucin produced by the normal gland and in the breast cancer-associated mucin, while others are exposed only in the latter. Detailed mapping of the amino acid sequence forming the core of the antibody-reactive epitopes has shown that in three cases, the epitope contains the single threonine found in the 20-amino acid tandem repeat sequence (Burchell et al., 1989). These results suggest that this threonine is not normally glycosylated, at least in some of the repeats.

The immunogenicity of both the normally processed and cancer-associated mucin in the mouse is well established, and mouse monoclonal antibodies directed to both core protein and carbohydrate epitopes are being widely used in diagnosis of breast and ovarian carcinomas. In addition to the B cell epitopes present in the PEM tandem repeat, it is now clear that T cell epitopes are also present and recognized by cytotoxic T cells derived from breast and pancreatic cancer patients (Barnd et al., 1989). The PEM molecule is therefore an extremely interesting and potentially useful antigen with repetitive B cell and T cell epitopes which can be specifically exposed in cancers. The fact that the mucin was found to contain a transmembrane domain, which was functional and capable of anchoring the molecule in the membrane of COS cells, indicates that this tumor-associated antigen is probably membrane-anchored and may be an appropriate target for both antibodies and T lymphocytes.

Acknowledgments—We thank Dr. I. Goldsmith for providing oligonucleotides and Kim Richardson for secretarial assistance.

Note Added in Proof—After submission of this study, a similar sequence for the same gene was reported (Ligtenberg, M. J. L., Gennissen, A. M. C., and Hilkens, J. (1989) J. Biol. Chem. 265, 5573-5578).

REFERENCES
Abe, M., Siddiqui, J., and Kufe, D. (1989) Biochem. Biophys. Res. Commun. 165, 644-649
Amerongen, A. V. N., Onderkerk, C. H., Roukema, P. A., Wolf, J. H., Lieshout, J. W., and Overdijk, B. (1983) Carbohydr. Res. 115, C1-C5
Andersson, S., Davis, D. L., Dahlbäck, H., Jornvall, H., and Russell, D. W. (1989) J. Biol. Chem. 264, 8222-8229
2 Kaufman, personal communication.
3 M. S. Lan, personal communication.
