Cloning, genomic organization, chromosomal assignment and expression of a novel mosaic serine proteinase: epitheliasin

Eric Jacquinet, Narayanam V. Rao, Gopna V. Rao, John R. Hoidal*

Department of Internal Medicine, Division of Respiratory, Critical Care and Occupational Medicine, Pulmonary Division, Wintrobe Building, Rm. 743A, 50N. Medical Drive, University of Utah Health Sciences Center and VA Medical Center, Salt Lake City, UT 84132, USA

Received 20 January 2000

Edited by Horst Feldmann

Abstract We report the isolation of a cDNA encoding a novel murine serine proteinase, epitheliasin. The cDNA spans 1753 bp and encodes a mosaic protein with a calculated molecular mass of 53 529 Da. Its domains include a cytoplasmic tail, a type II transmembrane domain, a low-density lipoprotein receptor class A domain, a cysteine rich scavenger receptor-like domain and a serine proteinase domain. The proteinase portion domain shows 46–53% identity with mouse neurotrypsin, acrosin, hepsin and enteropeptidase. The gene, located in the telomeric region in the long arm of mouse chromosome 16, consists of 14 exons and 13 introns and spans approximately 18 kb. Epitheliasin is expressed primarily in the apical surfaces of renal tubular and airway epithelial cells.

Key words: Serine proteinase; Mosaic protein; Epitheliasin

1. Introduction

Proteinases are implicated in a wide spectrum of physiologic and pathophysiological processes in the kidney. Renin, a proteinase synthesized in renal cortical cells plays a major role in the regulation of blood pressure and electrolyte balance by converting angiotensinogen to angiotensin I. Furthermore, the renal kallikrein-kinin system activated under conditions of mineralocorticoid excess represents a compensatory response against the development of hypertension and renal injury induced by salt excess. Proteolytic enzymes also have been ascribed important roles in both leukocyte-dependent and independent models of glomerular diseases (reviewed in [1]). Recently, Vallet and colleagues identified a novel serine proteinase from *Xenopus laevis* kidney epithelial cells, CAP 1, involved in activation of the epithelial sodium channel, EnaC [2]. This was the first report of channel activating activity of an endogenous proteinase.

In the present report, we describe a novel serine proteinase expressed in murine renal epithelial cells with sequence homology to CAP1. The enzyme, that we term epitheliasin, is a modular protein consisting of five sequence motifs, a cytoplasmic tail, a type II transmembrane (TM) domain, a low-density lipoprotein receptor class A (LDLRA)-like domain, a cysteine rich scavenger receptor-like (SRCR) domain and a serine proteinase domain. The sequence and structural features of epitheliasin cDNA and gene, its chromosomal localization and tissue expression are described. Epitheliasin has sequence identity to a human cDNA recently cloned by exon trapping named TMPRSS2 [3]. However, the tissue distribution of epitheliasin and TMPRSS2 is strikingly different.

2. Materials and methods

2.1. Materials

Multiple tissue Northern blots, ExpressHyb hybridization solution, rapid amplification of cDNA ends (RACE) ready cDNAs from mouse kidneys and Marathon cDNA kits were from CLONTECH (Palo Alto, CA, USA). TA cloning kits were from Invitrogen (Carlsbad, CA, USA). LA PCR kits were from Panvera (Madison, WI, USA). Klenow DNA polymerase, [α-^{32}P]dCTP (3000 Ci/mmol) and [γ-^{32}P]-dATP (3000 Ci/mmol) were from Amersham Life Science (Arlington Heights, IL, USA). BUPH™ Tris–glycine SDS, Tris–glycine and Immunogen Conjugation kits were from Pierce (Rockford, IL, USA). Alkaline phosphatase conjugated goat anti-rabbit antibody was from Zymed (San Francisco, CA, USA). BCIP/NBT tablets were from Sigma (St Louis, MO, USA). Citra solution and VIP substrate were from Vector Laboratories (Burlingame, CA, USA). Blocking reagent, SA-HRP and biotinyl tyramide were supplied by NEN Life Science Products (Boston, MA, USA).

2.2. Identification and cloning of epitheliasin cDNA

A conserved sequence around the serine active site residue (GGIDSCQGDSGGPLVC) was used to search the mouse EST database using TBLASTn. Of the 100 ESTs initially identified, a novel EST (ub58g01.s1) containing 389 nt and its mirror sequence (ub58g01.r1) were further analyzed using the non-redundant databases, BLASTn and BLASTx. Four overlapping sequences were found from these searches, one was from a kidney library (uc81c11.y1), two from a mammary gland library (vf86g09.r1, ve37e12.r1), and one from a blastocyst library (v164c03.r1).

To obtain the full-length cDNA of interest the RACE strategy was employed. Initially, LA PCR was utilized to amplify mouse kidney cDNA employing a sense primer (5′-$^{-36}$CCATACTGAACTCCTC-ATGCTGCT^{-13}-3′) designed based on the novel sequence and an anchor primer, AP1. The initial PCR product was subjected to nested PCR using a sense (5′-$^{-14}$CTGACACAGGCAGGATGGCATTG^{9}-3′) and an anti-sense primer (5′-^{1455}GTGGATTAGCTGTTCGCC-CTCATT^{1478}-3′). This nested reaction amplified a 1.5 kb product that was ligated into the pCR®3.1 vector and sequenced using an ABI automatic sequencer.

To obtain the 3′ end, mouse kidney cDNA was subjected to 3′-RACE. The cDNA was amplified using AP1 and a sense primer (5′-$^{-36}$CCATACTGAACTCCTCATGCTGCT^{-13}-3′). The product was diluted (1:50) and a nested PCR amplification was performed using a second anchor primer, AP2, and a sense primer (5′-$^{-14}$CTGACACA- GGCAGGATGGCATTG^{9}-3′). The 2 kb PCR product obtained was cloned and sequenced as described above.

2.3. Genomic cloning and analysis

To obtain the epitheliasin gene, a mouse genomic bacterial artificial chromosome (BAC) library (Genome Systems, St Louis, MO, USA) was screened using a 0.7 kb probe extending from 831 to 1477 nt of mouse epitheliasin cDNA. A single clone (BAC-24) was identified and confirmed by sequencing to contain the entire epitheliasin gene. To

*Corresponding author. Fax: (1)-801-585 3355.
E-mail: jhoidal@med.utah.edu

0014-5793/00/$20.00
PII: S0014-5793(00)01196-0

identify the intron junction borders, DNA from BAC-24 was directly sequenced using oligonucleotide primers defined initially by the cDNA sequences and subsequently by derived sequences. Southern analysis was used to determine the size of the epitheliasin gene.

2.4. Chromosomal assignment

The plasmid clone (BAC-24) obtained from the genomic library was used as a probe for chromosomal localization by fluorescence in situ hybridization (FISH). The probe was nick translation-labeled with biotin, hybridized to metaphase chromosomes and detected with Cy-3-conjugated streptavidin. Chromosome spreads were prepared by standard procedures and G-banded after trypsin treatment and Wright's staining. Hybridization and detection conditions on metaphase chromosomes were performed as previously described [4]. Probe signals were detected with the Cy3 conjugate viewed using an epifluorescence microscope. The fluorescence image was overlaid on the G-banded image to localize the gene.

2.5. Northern blot analysis

Mouse multi-tissue blots containing 2 μg of poly(A) RNA in each lane were prehybridized for 1 h at 68°C, then hybridized at 68°C with a 1.5 kb [α-^{32}P]dCTP-labeled probe that represented the coding region of the mouse epitheliasin cDNA. After low stringency washes, the blots were washed at high stringency at 50°C and autoradiographed.

2.6. Production of antibodies against epitheliasin

Rabbit polyclonal antiserum was raised to a synthetic peptide, cS^{331}HPNYDSKTKNND^{343}, located in the serine proteinase region of epitheliasin. The peptide was chosen based on predicted surface hydrophilicity and antigenicity. The peptide was coupled to keyhole-limpet hemocyanin. Subcutaneous injections were given to rabbits with 100 μg of conjugate that was emulsified in Freund's complete adjuvant and then boosted with the same amount of antigen in Freund's incomplete adjuvant at 2 week intervals until a titer of $>1{:}4000$ was obtained. The presence of anti-peptide antibodies was assessed by dot blot analysis using the peptide linked to ovalbumin as the antigen.

2.7. Immunohistology

Mouse kidneys and lungs were fixed in buffered 10% formaldehyde, and embedded in paraffin. Sections were cut at 5 μm depths, deparaffinized and rehydrated. Following antigen retrieval performed with 1× Citra solution in a microwave oven for 15 min at 700–900 W, the samples were washed in PBS. Endogenous peroxidase activity was blocked with 20% methanol and 3% H_2O_2 in PBS for 30 min at room temperature. The tissue was permeated using 10% Triton X-100 in PBS for 20 min at room temperature. Endogenous biotin was blocked by Vector Block avidin solution for 30 min at room temperature followed by Vector Blocking solution for 30 min at room temperature. The sections were then incubated with epitheliasin peptide anti-serum, dilution 1/500 in Block solution overnight at 4°C in a humid chamber. After washing with TNT, 1/500 horse anti-rabbit IgG serum in TNT was applied for 30 min at room temperature. The slides were then incubated with 1/100 SA-HRP in TNT for 30 min at room temperature. The signal was amplified with biotinyl tyramide for 5 min at room temperature. This was followed by a re-incubation with 1/100 SA-HRP in TNT. The signal was visualized using VIP substrate solution. The same process was applied to the slides used as controls, but epitheliasin anti-serum was replaced by non-immune rabbit serum.

3. Results and discussion

3.1. Cloning and analysis of the epitheliasin full-length cDNA

Fig. 1 shows the nucleic acid and deduced amino acid sequences of the complete cDNA reconstituted from the RACE fragments. As demonstrated by the immunohistochemistry described in a following section, the encoded protein is highly expressed in epithelial tissue. Accordingly, we named the protein epitheliasin. The composite cDNA spans 1753 nt. A 5′ untranslated region (UTR) extends 100 nt. The first in-frame ATG (1–3 nt) was assigned as the codon for the Met translation initiator since the sequence around this codon (AGGATGG) conforms to the Kozak consensus sequence for mammalian protein biosynthesis [5]. A single open reading frame begins with the ATG and extends 1470 nt. This is followed by a stop codon, TAA (1471–1473 nt) and a 3′-UTR of 152 nt, terminating in a poly(A)+tail of 28 nt. A consensus polyadenylation site (ATTAAA, 1600–1605 nt) is located 20 nt upstream of the poly (A)+tail.

3.2. Characteristics of the sequence and structural features of epitheliasin

The open reading frame encodes a protein of 490 amino acids with a calculated molecular mass of 53 529 kDa. Comparisons with sequences in GenBank, EMBL and SWISS-PROT reveal that the epitheliasin cDNA encodes a multidomain serine proteinase. A typical amino-terminal signal sequence is not present, but a hydrophobic region is present near the amino terminus (Leu^{84} to Trp^{105}). This 22 amino acid region is flanked by charged amino acids (Lys and Arg) and corresponds to a transmembrane domain [6]. Based on the difference in total charge between the 15-residue sequences on either side of the membrane-spanning domain epitheliasin can be classified as a type II integral membrane-bound protein [7,8] that has a cytosol facing amino-terminal tail region consisting of 83 amino acids (Met^{1} to Ser^{83}) and an extracellular facing COOH-terminal modular region. The absence of a signal peptide and the presence of a transmembrane domain in epitheliasin are analogous to homologous serine proteinases, enteropeptidase, a key enzyme in digestion that is responsible for the conversion of trypsinogen to trypsin [9], hepsin, a membrane-associated proteinase involved in the formation of thrombin on cell surfaces [10], and a recently described human airway trypsin-like proteinase [11].

The predicted domain structure of epitheliasin is shown in Fig. 2. A LDLRA domain extending from Cys^{112} to Cys^{147} and containing six cysteines follows the transmembrane domain. This domain motif is found in a number of proteins that are functionally unrelated to the LDLR family, including clotting proteinases and enteropeptidase. In each of these proteins the domain is thought to function as a protein-binding domain. The LDLRA domain in epitheliasin is similar to other typical LDLRA domains that are about 40 amino acids long and contain six cysteines [12]. The cysteines form intra-domain bridges resulting in a cluster of negatively charged residues in a single loop positioned for high affinity binding to positively charged sequences in LDLR ligands.

Following the LDLRA domain, an SRCR-like domain extends from va1^{48} to Gly^{243}. SRCR domains are classified into two groups, group A and B according to the number of conserved cysteine residues, six or eight, respectively [13]. In a recent analysis, all but one of the 33 independent SRCR domains that had been previously identified had six or eight cysteines [14]. An unusual feature of this domain in epitheliasin is that it contains only four cysteines. These cysteine residues in epitheliasin are completely conserved in positions suggesting that the domain belongs to group A. The SRCR domain that is closest to that in epitheliasin is present in complement factor 1 (CF1), a serum proteinase that regulates the complement cascade by cleaving C3b and C4b. CF1 contains a single SRCR domain with five cysteines [13].

The function of SRCR domains is largely unknown. It seems likely that most of these domains are involved in bind-

```
-100 GAAGCCAAGA GCTCGGACAG AGGCGGAGAG GGGCGACAAC GCAACAGGTC AACTACAGGA AGCCCCATAC TGAACTCCTC ATGCTGCTGA CACAGGCAGG ATG GCA TTG AAC TCA GGG    18
                                                                                                                  M   A   L   N   S   G    6

 19 TCA CCT CCA GGA ATC GGA CCT TGC TAT GAG AAC CAC GGG TAT CAG TCT GAG CAC ATC TGT CCT CCG AGA CCA CCA GTG GCT CCC AAT GGC TAC AAC TTG TAT   120
     S   P   P   G   I   G   P   C   Y   E   N   H   G   Y   Q   S   E   H   I   C   P   P   R   P   P   V   A   P   N   G   Y   N   L   Y  40

121 CCA GCC CAG TAC TAC CCA TCT CCA GTG CCT CAG TAT GCT CCG AGG ATT ACA ACG CAA GCC TCA ACA TCT GTC ATC CAC ACA CAT CCC AAG TCC TCA GGA GCA   222
     P   A   Q   Y   Y   P   S   P   V   P   Q   Y   A   P   R   I   T   T   Q   A   S   T   S   V   I   H   T   H   P   K   S   S   G   A  74

223 CCG TGC ACC TCA AAG TCT AAG AAA TCG CTG TGT TTA GCC CTT GCC CTG GGC ACT GTC CTC ACG GGA GCT GCT GTG GCT GCT GTC TTG CTT TGG AGG TTC TGG   324
     P   C   T   S   K   S   K   K   S   L   C   L   A   L   A   L   G   T   V   L   T   G   A   A   V   A   A   V   L   L   W   R   F   W  108

325 GAC AGC AAC TGT TCT ACG TCT GAG ATG GAG TGT GGG TCT CTA GGC ACA TGC ATC AGC TCT TCT CTC TGG TGT GAC GGG GTA GCA CAT TGT CCC AAC GGA GAA   426
     D   S  [N   C   S] T   S   E   M   E   C   G   S   L   G   T   C   I   S   S   S   L   W   C   D   G   V   A   H   C   P   N   G   E  142

427 GAT GAG AAC CGT TGT GTT CGT CTC TAC GGA CAA AGC TTC ATC CTC CAG GTT TAC TCA TCT CAG AGG AAA GCC TGG TAT CCC GTG TGC CAG GAT GAT TGG AGT   528
     D   E   N   R   C   V   R   L   Y   G   Q   S   F   I   L   Q   V   Y   S   S   Q   R   K   A   W   Y   P   V   C   Q   D   D   W   S  176

529 GAG AAC TAC GGG AGA GCA GCA TGT AAA GAC ATG GGA TAC AAG AAC AAT TTT TAT TCC AGC CAA GGG ATA CCA GAC CAG AGC GGG GCA ACG AGC TTT ATG AAG   630
     E   N   Y   G   R   A   A   C   K   D   M   G   Y   K   N   N   F   Y   S   S   Q   G   I   P   D   Q   S   G   A   T   S   F   M   K  210

631 CTG AAT GTG AGC TCA GGC AAT GTT GAC CTC TAT AAA AAA CTC TAC CAC AGT GAC TCA TGT TCA TCC CGC ATG GTG GTT TCT TTG CGC TGT ATA GAA TGC GGG   732
     L  [N   V   S] S   G   N   V   D   L   Y   K   K   L   Y   H   S   D   S   C   S   S   R   M   V   V   S   L   R   C   I   E   C   G  244

733 GTT CGC TCA GTG AAA CGC CAG AGC AGG ATT GTG GGT GGA TTG AAT GCC TCA CCA GGA GAC TGG CCC TGG CAG GTC AGC CTG CAC GTC CAA GGC GTC CAC GTC   834
     V   R   S   V   K   R   Q   S   R   I   V   G   G   L   N   A   S   P   G   D   W   P   W   Q   V   S   L   H   V   Q   G   V   H   V  278

835 TGC GGA GGC TCC ATC ATC ACC CCC GAG TGG ATT GTG ACG GCC GCC CAC TGT GTG GAA GAA CCC CTC AGC GGC CCG AGG TAC TGG ACG GCA TTT GCG GGA ATT   936
     C   G   G   S   I   I   T   P   E   W   I   V   T   A   A  (H)  C   V   E   E   P   L   S   G   P   R   Y   W   T   A   F   A   G   I  312

937 CTG AGA CAG TCT CTC ATG TTC TAT GGA AGT AGA CAC CAG GTA GAA AAA GTA ATT TCC CAT CCA AAT TAC GAC TCT AAG ACC AAG AAT AAC GAC ATT GCT CTC   1038
     L   R   Q   S   L   M   F   Y   G   S   R   H   Q   V   E   K   V   I   S   H   P   N   Y   D   S   K   T   K   N   N  (D)  I   A   L  346

1039 ATG AAG CTG CAG ACA CCT TTG GCT TTT AAT GAT CTA GTG AAG CCA GTG TGT CTG CCG AAC CCA GGC ATG ATG CTA GAC CTA GAC CAG GAA TGC TGG ATT TCG   1140
      M   K   L   Q   T   P   L   A   F   N   D   L   V   K   P   V   C   L   P   N   P   G   M   M   L   D   L   D   Q   E   C   W   I   S  380

1141 GGG TGG GGG GCC ACC TAT GAG AAA GGG AAG ACC TCG GAC GTG TTG AAT GCT GCC ATG GTA CCC TTG ATC GAG CCC TCC AAA TGT AAT AGT AAA TAC ATA TAC   1242
      G   W   G   A   T   Y   E   K   G   K   T   S   D   V   L   N   A   A   M   V   P   L   I   E   P   S   K   C   N   S   K   Y   I   Y  414

1243 AAC AAC CTA ATC ACA CCA GCC ATG ATC TGT GCC GGC TTC CTC CAG GGG TCT GTC GAC TCT TGC CAG GGA GAC AGT GGA GGG CCG CTG GTT ACT TTG AAG AAT   1344
      N   N   L   I   T   P   A   M   I   C   A   G   F   L   Q   G   S   V   D   S   C   Q   G   D  (S)  G   G   P   L   V   T   L   K   N  448

1345 GGG ATC TGG TGG CTG ATT GGG GAC ACG AGC TGG GGC TCG GGC TGT GCC AAG GCA CTC AGA CCT GGA GTA TAC GGG AAC GTG ACG GTA TTT ACA GAT TGG ATC   1446
      G   I   W   W   L   I   G   D   T   S   W   G   S   G   C   A   K   A   L   R   P   G   V   Y   G  [N   V   T] V   F   T   D   W   I  482

1447 TAC CAG CAA ATG AGG GCG AAC AGC TAA TCCACATGGC TTTGTCCCAG ACTTCCTTTG TCTTCAACAA CCTTTTGCAA GAAAACCAAG GGCCTGAATT TTAACTTCCT GTGCACAATG   1563
      Y   Q   Q   M   R   A   N   S   *                                                                                                  490

1564 TACCTTTTGA GATGATTCGA AGGGCCTTTC ACTTTTATTA AACAGTGACT TGTTTGACTG TGAAAAAAAA AAAAAAAAAA AAAAAAAAAA                                   1653
```

Fig. 1. Nucleic acid and deduced amino acid sequences of full-length cDNA encoding epitheliasin. Nucleic acid numbering starts at the A nt of the putative translation initiation codon (ATG), with positive and negative numbers proceeding to downstream and upstream of sequence, respectively. The consensus translation initiation codon encompassed by ribosomal binding site sequence and putative polyadenylation signal are underlined. An asterisk shows the termination codon. Amino acid residues in single letter code are numbered starting at the first Met residue in the open reading frame and the numbers are shown on the right end of each line. Potential *N*-linked glycosylation sites are boxed and the encircled amino acid residues are those of the catalytic triad.

ing to molecules on the cell surface or in the extracellular space. Direct evidence supporting the idea that SRCR domains mediate binding to other cell surface proteins or extracellular proteins has recently been provided [14,15].

3.3. Features of serine proteinase domain

The proteinase domain begins with Ile^{254} and represents the major domain (about 50%) of the encoded protein. The predicted molecular mass of the domain is 25 892 kDa. The domain contains all the major features common to the S1 family of the chymotrypsin (or SA) clan of serine proteinases. The residues contributing to the salient structural features in chymotrypsin include: (1) His^{57}, Asp^{102}, and Ser^{195} that make up the catalytic triad, (2) Gly^{193}, Asp^{194} and Ser^{195} that form an oxyanion hole required for catalytic efficiency, (3) Ser^{214}, Trp^{215} and Gly^{216} that bind the main-chain of a substrate, and (4) residues that occupy the bottom (Ser^{189}) and sides (Gly^{216} and Gly^{226}) of the substrate specificity pocket (S_1 subsite). All of the residues contributing to the first three features and the residues Gly^{216} and Gly^{226} on the sides of the substrate specificity pocket of chymotrypsin are strictly conserved in epitheliasin. However, in epitheliasin the residue corre-

Table 1
Exon–intron junctions organization of epitheliasin gene

3' splice site	Exon size in Amino acid	5' splice site	Phase
		-59 CAA CAG *GT*GAGAAGCGCGCCG	
GTTTCTTTCCTTC*AG* GTC ACC-44	..... (5)	AAC TCA N^4 S^5 *GT*AAGTGCTAATTCT	0
TTTCCCATTGTTT*AG*	GGG TCA ..... G^6 S^7 (73)	ACC TCA A T^{77} S^{78} *GT*AAGACTCCTTAGC	0
CTTTTCTTCCCGC*AG*	AG TCT ..... K^{79} S^{80} (29)	AGG TTC T R^{106} F^{107} *GT*AAGTTGGGGGCTG	I
CCAATACAATGCC*AG*	GG GAC ..... W^{108} D^{109} (39)	AAC CG N^{145} R^{146} *GT*GTTGTGAGTTATC	I
TTCTTTCTCCTTC*AG* T	TGT GTT ..... C^{147} V^{148} (44)	TAC AA Y^{189} K^{190} *GT*GAGTATGGAAGCC	I
TGTCTTTTTTTCC*AG* G	AAC AAT ..... N^{191} N^{192} (37)	CAC AG H^{226} S^{227} *GT*ATGGAGTTTTTTC	II
CTTTTTTCTTTCC*AG* T	GAC TCA ..... D^{228} S^{229} (14)	TGT ATA G C^{240} I^{241} *GT*GAGTGAGTACTTC	II
GCTTGTCACCCTC*AG*	AA TGC ..... E^{242} C^{243} (57)	GAA GA E^{297} E^{298} *GT*ATGCCTCCATTCT	I
CTTCTGTCTCTCA*AG* A	CCC CTC ..... P^{299} L^{300} (58)	TTT AAT G F^{355} N^{356} *GT*ACGTGAGACTCAG	II
CTCTTCTTTAAAC*AG*	AT CTA ..... D^{357} L^{358} (32)	GAG AAA G E^{387} K^{388} *GT*GAGGCTTTGGGTC	I
TGCCTCTGTTGTT*AG*	GG AAG ..... G^{389} K^{390} (48)	TGC CAG C^{435} Q^{436} *GT*AATTGTGAGTGGT	I
TGCTGTGTTCCCC*AG*	GGA GAC ..... G^{437} D^{438} (51)	ATG AGG M^{486} R^{487} *GT*TATTTCCTCTATT	0
TTCCTATTTGCAC*AG*	GCG AAC ..... A^{488} N^{489} (3)		0

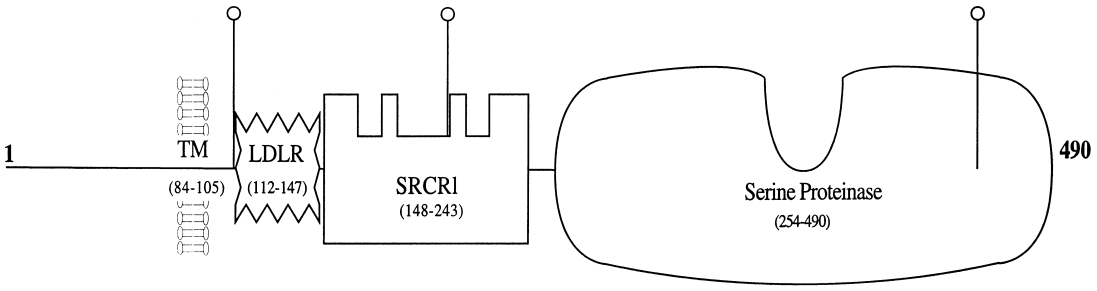

Fig. 2. The domain organization of epitheliasin. Starting at the NH_2-terminus the epitheliasin contains a TM domain followed by a LDLRA domain, a SRCR domain, and finally the serine proteinase domain. *N*-glycosylation sites are indicated by a circle. The numbers in parentheses refer to the amino acid residues of each domain.

sponding to Ser^{189} of chymotrypsin is replaced by an acidic residue, Asp. This suggests that epitheliasin has specificity for cleavage after Lys or Arg, indicating a trypsin-like substrate specificity for the enzyme.

Comparison of the amino acid sequence encoding the proteinase domain in epitheliasin with other serine proteinases indicates that this region of the protein shares identity with mouse enteropeptidase (53%), hepsin (51%), acrosin (48%), and neurotrypsin (46%), all multi-domain members of the chymotrypsin family of serine proteinases with trypsin-like substrate specificity. The aforementioned CAP 1 from *Xenopus laevis* kidney epithelial cells has a sequence identity with epitheliasin of 44%.

Based on findings with related vertebrate trypsinogens we predict that epitheliasin is synthesized as an inactive zymogen that is converted to an active serine proteinase by cleavage of the Arg^{253}–Ile^{254} peptide bond in the extracellular domain of the enzyme. Most vertebrate trypsinogens are activated by proteolytic cleavage of a Lys (Arg)–Ile bond. The identity or the origin of the proteinase responsible for this cleavage in epitheliasin is not known. One possibility is that epitheliasin is synthesized as a single-chain zymogen and undergoes intracellular cleavage and activation by a furin-like enzyme prior to insertion into the membrane. This is based on the Arg-Gln-Ser-Arg^{253} sequence that immediately precedes the Ile-Val-Gly-Gly^{257} representing the NH_2-terminus of the proteinase domain. Arg-X-X-Arg motifs are furin recognition sequences [16–20]. Interestingly, all the domains of epitheliasin are flanked by recognition sites for furin-like enzymes, suggesting the need to clarify the role of furin-like enzymes in processing of epitheliasin.

Based on the structure of enteropeptidase and a comparison with other chymotrypsin-like serine proteinases, we also predict that epitheliasin, following intracellular cleavage, forms two chains with the smaller chain containing the proteinase domain, and the larger the membrane-spanning segment, and the LDLRA and SRCR-like domains that may serve as substrate recognition sites. Several chymotrypsin-like serine proteinases including enteropeptidase have a disulfide bond that covalently links the two chains [21]. The proteinase domain in epitheliasin contains eight Cys residues in conserved positions. By comparison with chymotrypsin, three of the Cys pairs (42/58, 168/182 and 191/220) that form disulfide bond loops around His^{57}, Met^{180} and Ser^{195} are conserved in epitheliasin. Although the other two cysteines (Cys^{122} and Cys^{136}) are located in conserved positions, their pairing counterparts Cys^{1} and Cys^{201} that are involved in interchain disulfide bonds are absent. This suggests that epitheliasin is likely distinct from enteropeptidase and other multidomain serine proteinases in that it lacks disulfide bond(s) between the proteinase motif and the rest of the protein [22]. Thus, the mechanism of association of the two chains in epitheliasin is not clear.

Three asparagine-linked glycosylation sites are present in epitheliasin, Asn^{111} located at the beginning of the LDLRA domain of the protein, Asn^{212} located in the SRCR domain and Asn^{474} located in the proteinase domain (see Fig. 1). Other features of the deduced primary structure of the protein include a cAMP- or cGMP-dependent protein kinase phosphorylation site (Lys^{249}–Ser^{252}). Two protein kinase C phosphorylation sites are present in the cytoplasmic domain (Thr^{77}–Lys^{79} and Thr^{80}–Lys^{82}), three in the SRCR domain (Ser^{162}–Arg^{164}, Ser^{231}–Arg^{233}, Ser^{237}–Arg^{249}), one between the SRCR domain and the proteinase domain (Ser^{257}–Lys^{249}), and one in the proteinase domain (Thr^{445}–Lys^{447}). Three casein kinase II phosphorylation sites are present, two in the LDLRA domain (Ser^{113}–Glu^{116}, Ser^{116}–Glu^{119}), and the last one in the proteinase domain (Ser^{261}–Asp^{264}). Finally, an ATP/GTP-binding site motif A is present in the proteinase domain of epitheliasin, from Ile^{379} to Ala^{396}. This motif is found in a number of proteins including those in the myosin and Ras families. The relevance of these various sites in epitheliasin is not presently known.

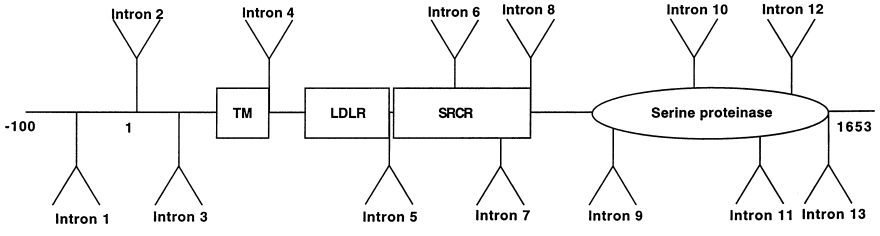

Fig. 3. Schematic representation of the genomic organization of epitheliasin. The intron placements are depicted in relationship to the domains of the mouse epitheliasin protein. The numbering represents nucleotides.

3.4. Genomic organization

The epitheliasin gene contains 14 exons separated by 13 introns (Fig. 3). The first exon is located in the 5′ untranslated region. The last exon contains 9 bp of the coding sequence, the stop codon and the 3′ untranslated region. The exon distribution reflects the organization of the deduced protein. Exon 2 and 3, respectively 68 and 220 nt (M^{1}–S^{78}), encode for the cytoplasmic domain. Exon 4, 87 nt, (K^{79}–F^{107}) encodes for the transmembrane domain. Exon 5, 117 nt, (D^{109}–R^{146}) encodes for the LDLR domain (C^{112}–C^{147}). An unusual feature of epitheliasin is that the SRCR domain is encoded by three exons, 6–8, respectively 130 nt, 111 and 44 nt (C^{147}–I^{241}). Usually SRCR domains are encoded by one or two exons, in regard to type B or type A, respectively. Exons 9–13, respectively 169, 176, 96, 143 and 153 nt, (E^{242}–R^{490}) encode for the serine protease domain. Vertebrate serine protease-like genes have been grouped into five classes based on intron positions [23]. The gene organization of the epitheliasin protease domain is typical of second group containing members of the trypsin family of serine proteases and consisting of five exons with each of the three components of the catalytic triad encoded by sequences in a different exon. In epitheliasin, the catalytic histidine is located in exon 9, the aspartic in exon 10 and the serine in exon 13. In general, the organization of epitheliasin is similar to that of other multiple domain serine proteinases. Each domain is coded in an independent manner by one or more exons. A common feature among all multidomain protease cloned to date is the five exons coding for the serine proteinase domain [24].

As shown in Table 1, all intron/exon junctions contain the expected GT splice donor and AG splice acceptor sites and conform to the consensus sequences established for intronic donor and acceptor splice signals [25]. Four introns are inserted between codons (type 0 splice junction), five are after the first nucleotide in a codon (type I splice junction), and four after the second nucleotide codon (type II splice junction). six bands were strongly positive by Southern analysis with sizes of 7000, 5000, 2700, 1400, 1200 and 900 nt. Adding the size of the fragments indicates that the epitheliasin gene is approximately 18 kb.

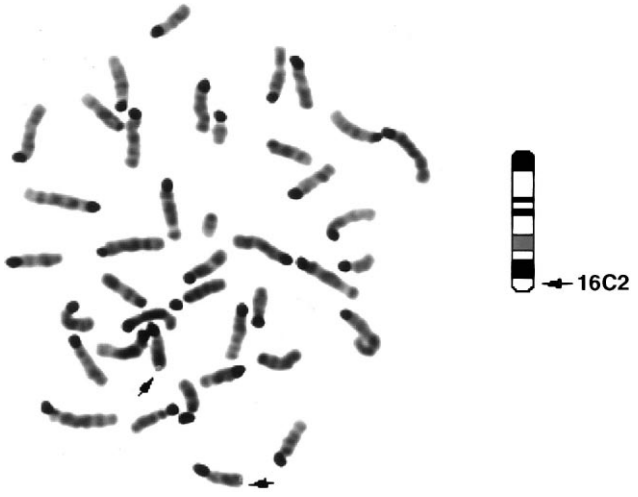

Fig. 4. In situ hybridization of a biotin-labeled epitheliasin probe to mouse metaphase cells. The chromosome 16 homologues are identified with arrows. Specific labeling was observed at chromosome band 16C2.

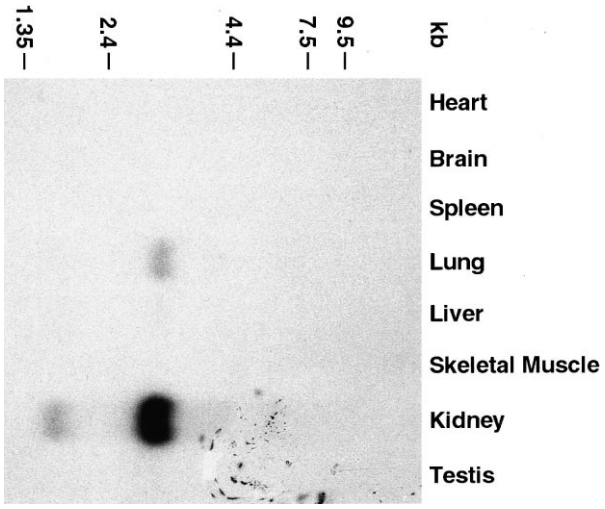

Fig. 5. Northern blot analysis of epitheliasin mRNA in various mouse tissues. Each lane contained 2 μg of poly(A)+RNA. The blot was hybridized to an epitheliasin cDNA probe.

3.5. Chromosomal assignment

FISH was performed on normal mouse chromosomes using a BAC containing the genomic sequence of epitheliasin (Fig. 4). These studies localized the epitheliasin gene to the telomeric region in the long arm of chromosome 16. The band localization was confirmed on G-banded chromosomes. The hybridization efficiency was 92.5%. No other serine proteinases have been localized to this region. The region is homologous with the so-called 'Down's syndrome region' of human chromosome region 21q22.2 and 21q22.3.

3.6. Expression of epitheliasin mRNA in vivo

The in vivo distribution of epitheliasin mRNA was investigated in adult mouse tissues by Northern blot analysis. As shown in Fig. 5, a prominent 2.8 kb transcript and a less prominent 1.5 kb transcript were observed in the kidney. Because of preliminary results that suggest an alternative polyadenylation site approximately 1.3 kb downstream from the initial polyadenylation site, we believe that the weaker signal actually represents the characterized cDNA. A prominent 2.8 kb signal was also seen in the lung and a weaker signal of similar size was observed in liver tissue. No signal was observed in heart, brain, spleen, testis or skeletal muscle. Of note, all tissues that express epitheliasin have epithelial cells as a prominent feature of their cellular makeup.

3.7. Immunohistochemical localization

Fig. 6A shows the kidney in which only tubular epithelial cells are stained with no staining of glomeruli. The staining is restricted to cells located in distal tubules. The staining is most intense at the apical pole of the cells, facing the lumen of the tubules. The staining is faint in the cytoplasm, basal and lateral side of the cells. Fig. 6B shows the lung in which staining is primarily limited to the apical surface of airway epithelial cells. Staining is minimal or absent in the vasculature and alveolar spaces. No staining was observed in control slides. Further analysis by in situ hybridization using a 300 nt epitheliasin riboprobe demonstrated that the pattern of gene expression was the same as that of protein expression (data

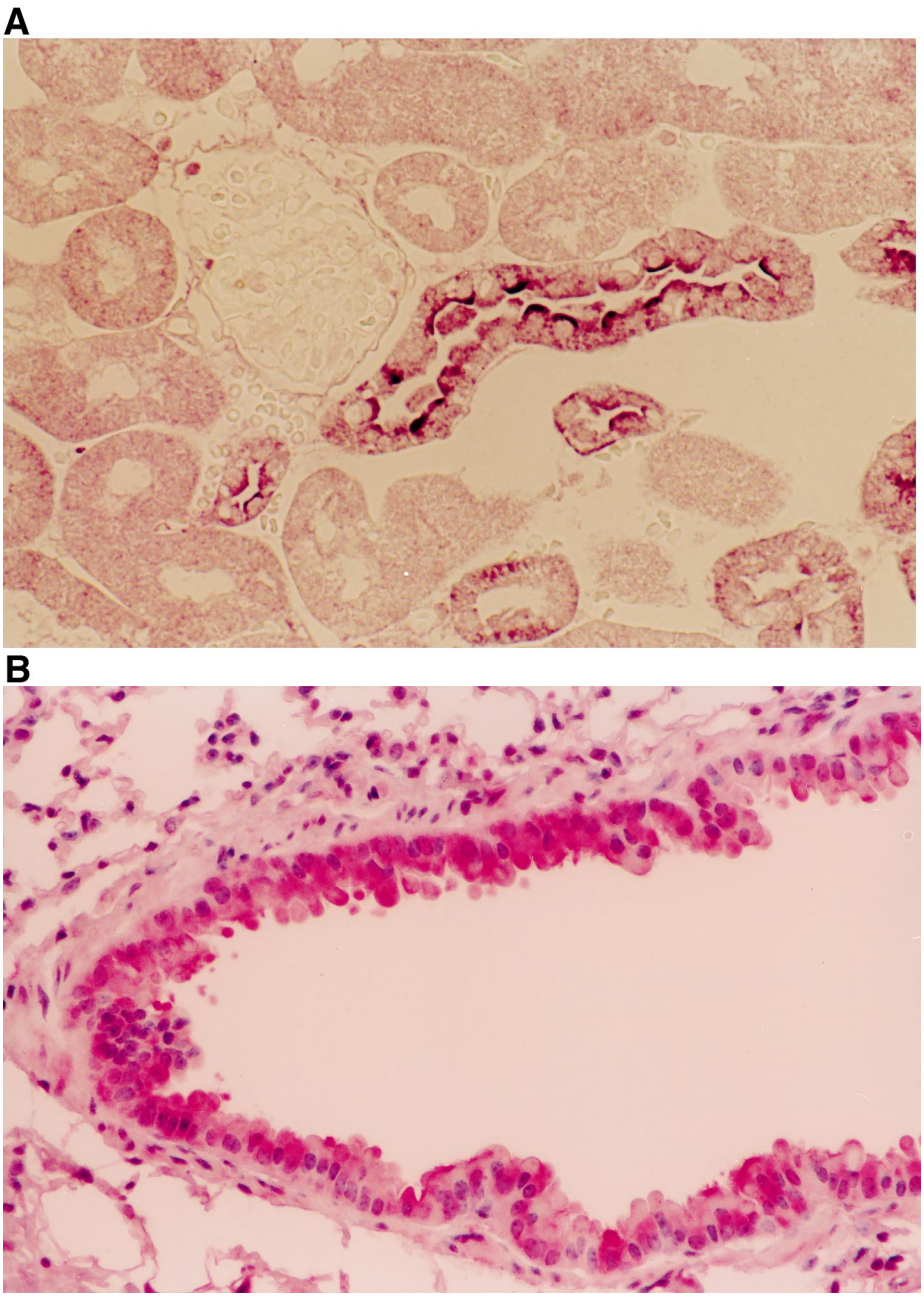

Fig. 6. Immunohistochemical localization of epitheliasin in adult mouse tissue. A: A section from the kidney (magnification 20×). Positive staining is seen in apical region of renal distal tubule epithelial cells. B: A section from lung (magnification 20×). Positive staining is seen in bronchial epithelial cells. No stain was observed in control sections in which normal rabbit serum substituted for rabbit anti-mouse epitheliasin (data not shown).

not shown). These results support the epithelial and membrane localization of epitheliasin.

During the course of this investigation Paolini-Giacobino and colleagues reported on a human cDNA cloned by exon trapping named TMPRSS2 [3]. The portion of the TMPRSS2 cDNA that was reported has approximately 80% sequence identity to epitheliasin. However, the tissue distribution of epitheliasin and TMPRSS2 is strikingly different. While epitheliasin is highly expressed in the mouse kidney, no expression of TMPRSS2 was observed in the human kidney. In contrast, no expression of epitheliasin was observed in the mouse heart or brain, while a high level of expression of TMPRSS2 was observed in human heart and an intermediate level in brain. Moreover, the size of epitheliasin of the mRNA transcript (2.8 kb) and that of TMPRSS2 (3.8 kb) are different. Whether TMPRSS2 is the human orthologue of epitheliasin or a closely related gene product will require further study.

The biological role of epitheliasin is not known. The homology with CAP1 and apical membrane distribution raise the possibility that epitheliasin may activate ion transport channels of the plasma membrane. In addition, cell-surface proteinases of normal and malignant cells are thought to play roles in cell growth, chemotaxis, endocytosis, exocytosis,

blood coagulation, fibrinolysis and tissue invasion during metastasis [26]. While the function of the non-proteinase domains is unexplored, the presence of these domains with a modular organization represents a common feature of regulatory serine proteinases (e.g. proteinases of the fibrinolytic and blood coagulation systems). Studies of the kinetic effects of deleting the non-proteinase domain from enteropeptidase clearly implicate it in the recognition of macromolecular substrates and inhibitors [21].

Acknowledgements: The work was supported by HL 50153 and HL 37615 from the NHLBI. We gratefully acknowledge the assistance of Dr. Kurt Albertine and Zhengming Wang for the immunohistochemical studies. GenBank accession number for epitheliasin nucleotide sequence: BankIt243070 AF113596.

References

[1] Baricos, W.H. and Shah, S.V. (1991) Kidney Int. 40, 161–173.
[2] Vallet, V., Chraibi, A., Gaeggeler, H.P., Horisberger, J.D. and Rossier, B.C. (1997) Nature 389, 607–610.
[3] Paoloni-Giacobino, A., Chen, H., Peitsch, M.C., Rossier, C. and Antonarakis, S.E. (1997) Genomics 44, 309–320.
[4] Pinkel, D., Straume, T. and Gray, J.W. (1986) Proc. Natl. Acad. Sci. USA 83, 2934–2938.
[5] Kozak, M. (1986) Cell 44, 283–292.
[6] von Heijne, G. and Manoil, C. (1990) Protein Eng. 4, 109–112.
[7] High, S. (1992) BioEssays 14, 535–540.
[8] Semenza, G. (1986) Annu. Rev. Cell Biol. 2, 255–313.
[9] Matsushima, M., Ichinose, M., Yahagi, N., Kakei, N., Tsukada, S., Miki, K., Kurokawa, K., Tashiro, K., Shiokawa, K., Shinomiya, K., Umeyama, H., Inoue, H., Tatahashi, T. and Takahashi, K. (1994) J. Biol. Chem. 269, 19976–19982.
[10] Kazama, Y., Hamamoto, T., Foster, D.C. and Kisiel, W. (1995) J. Biol. Chem. 270, 66–72.
[11] Yamaoka, K., Masuda, K.-I., Ogawa, H., Takagi, K.-I., Umemoto, N. and Yasuoka, S. (1998) J. Biol. Chem. 273, 11895–11901.
[12] Sudhof, T.C., Goldstein, J.L., Brown, M.S. and Russell, D.W. (1985) Science 228, 815–822.
[13] Resnick, D., Pearson, A. and Krieger, M. (1994) Trends Biochem. Sci. 19, 5–8.
[14] Whitney, G.S., Starling, G.C., Bowen, M.A., Modrell, B., Siadak, A.W. and Aruffo, A. (1995) J. Biol. Chem. 270, 18187–18190.
[15] Bowman, A. and Drummond, A.H. (1984) Br. J. Pharmacol. 81 (4), 665–674.
[16] Bresnahan, P.A., Hayflick, J.S., Molloy, S.S., and Thomas, G. (1993) in: Mechanisms of Intracellular Trafficking and Processing of Proproteins (Loh, Y.P., Ed.), pp. 225–250, CRC Press, Boca Raton, FL.
[17] Creemers, J.W.M., Siezen, R.J., Roebroek, A.J.M., Ayoubi, T.A.Y., Huylebroeck, D. and Van de Ven, W.J.M. (1993) J. Biol. Chem. 268, 21826–21834.
[18] Hatsuzawa, K., Nagahama, M., Takahashi, S., Takada, K., Murakami, K. and Nakayama, K. (1992) J. Biol. Chem. 267, 16094–16099.
[19] Molloy, S.S., Bresnahan, P.A., Leppla, S.H., Klimpel, K.R. and Thomas, G. (1992) J. Biol. Chem. 267, 16396–16402.
[20] Van de Ven, W.J.M. and Roebroek, A.J.M. (1993) Crit. Rev. Oncog. 4, 115–136.
[21] Lu, D., Yuan, X. and Xinglong Z, S.J.E. (1997) J. Biol. Chem. 272 (50), 31293–31300.
[22] Delabar, J.M., Tehophile, D., Rahmani, Z., Chettouh, Z., Blouin, J.L., Prieur, M., Noel, B. and Sinet, P.M. (1993) Eur. J. Hum. Genet. 1, 114–124.
[23] Irwin, D.M., Robertson, K.A. and MacGillivray, R.T. (1988) J. Mol. Biol. 200, 31–45.
[24] Cool, D.E. and MacGillivray, R.T. (1987) J. Biol. Chem. 262, 13662–13673.
[25] Breathnach, R. and Chambron, P. (1981) Annu. Rev. Biochem. 50, 349–383.
[26] Bond, J.S. (1991) Biomed. Biochim. Acta 50, 775–780.