Biochemical and Biophysical Research Communications **290,** 452–456 (2002)
doi:10.1006/bbrc.2001.6224, available online at http://www.idealibrary.com on IDEAL®

Identification of a Novel Protease Inhibitor Gene That Is Highly Expressed in the Prostate

Åke Lundwall[1] and Adam Clauss
Department of Laboratory Medicine, Lund University, University Hospital MAS, Malmö, Sweden

Received December 6, 2001

A novel gene was identified 52 kb upstream of the gene encoding the protease inhibitor elafin (*PI3*) on human chromosome 20q12-13.1. The transcript of the new gene, denoted *huWAP2,* was characterized by rapid amplification of cDNA ends and DNA sequencing. The size is 774 bp and it gives rise to a polypeptide of 111 amino acid residues that is homologous to elafin and similar WAP-type protease inhibitors. By RT-PCR it was shown that the gene is highly expressed in prostate, skin, lung, and esophagus. © 2002 Elsevier Science

Key Words: **semen; seminal plasma; protease inhibitor; WAP domain; prostate.**

The mammalian seminal vesicles secrete a limited number of proteins at very high concentration. The primary structures of these proteins display a remarkably high species variation. Analyses of their genes show that this is caused by a rapid and unusual evolution of a single exon containing most of the translated nucleotides (1). In contrast, the first and last exons, encompassing signal peptide-coding nucleotides and 3′ non-translated nucleotides, are conserved between species and appears to evolve at a normal rate.

In human, there are two predominant seminal vesicle secreted proteins, semenogelin I and semenogelin II, which are encoded by the genes *SEMG1* and *SEMG2,* located at chromosome 20q12-13.1 (2–4). Using the conserved nucleotide sequence of exon 1 and exon 3 it was possible to identify two related human genes (5). These codes for the two protease inhibitors elafin and secretory leukocyte protease inhibitor (SLPI). In the elafin gene, the second exon codes for the protease inhibitor domain and a region containing 4 repeats that can serve as substrate for transglutaminase (6). In the SLPI gene, the exon encoding the protease inhibitor domain has been duplicated so that the gene contain 4 exons (7). The structure of the protease inhibitor domain contains 4 disulfides and because of this, it is known as a four-disulfide core-domain. The first example of a protein with this structure was a predominant protein found in milk whey of lactating mouse. The protein of 14 kDa is known as whey acidic protein (WAP) and because of this, the domain is also known as a WAP domain (8).

In a previous study, it was shown that the gene encoding elafin, *PI3,* is located to the same chromosomal region as *SEMG1* and *SEMG2* (9). During the accumulation of nucleotide sequence data by the human genome project, it has also become evident that these three genes as well as *SLPI,* are closely linked within a region of 75 kb on chromosome 20. To identify related genes, DNA located in the vicinity of the *SEMG1* locus was surveyed for nucleotide sequences with similarity to the first exon of *SEMG1, SEMG2, PI3,* and *SLPI.*

MATERIAL AND METHODS

RT-PCR. RNA samples were isolated from tissue specimens homogenized in 4 M guanidinium thiocyanate as described (10). Tissue samples from the urogenital tract and mammary glands were from patients undergoing surgical treatment for neoplastic disease while other specimens were taken at autopsy approximately 20 h postmortem. The Helsinki Declaration regarding the use of human tissues was followed. RNA from human lung, pancreas, salivary gland, skeletal muscle and trachea were purchased from Clontech. Oligo-dT-primed cDNA synthesis was done with 3 μg of total RNA in a volume of 15 μl using the First-Strand cDNA synthesis kit (Amersham–Pharmacia). The subsequent PCR was done with 2 μl cDNA, equivalent to approximately 0.4 μg of RNA, in a volume of 10 μl using the Advatage 2 PCR kit (Clontech) and 0.2 μM of gene-specific primers. The transcript of the housekeeping gene, adenine phosphoribosyltransferase (APRT), served as an internal control. The two new gene primers, TTGGTCCTCATGGTGTCTCTCGTT (N1F1) and GCCACAGTGCAGGTAACAACACTT (N1R1), and the two APRT primers, GCCGCATCGACTACATCGCAGGCCTAGA and CTCACAGGCAGCGTTCATGGTTCCACCA, were purchased from Life Technologies (UK). The PCRs were run on a MJ Research PTC-200 with a program consisting of an initial denaturation at 95°C for 1 min, followed by 35

The novel nucleotide sequence data published here have been deposited with the GenBank sequence databank and are available under Accession No. AY037803.

[1] To whom correspondence and reprint requests should be addressed at Department of Laboratory Medicine, University Hospital MAS, S-205 02 Malmö, Sweden. Fax: +46-40337043. E-mail: Ake.Lundwall@klkemi.mas.lu.se.

0006-291X/02 $35.00

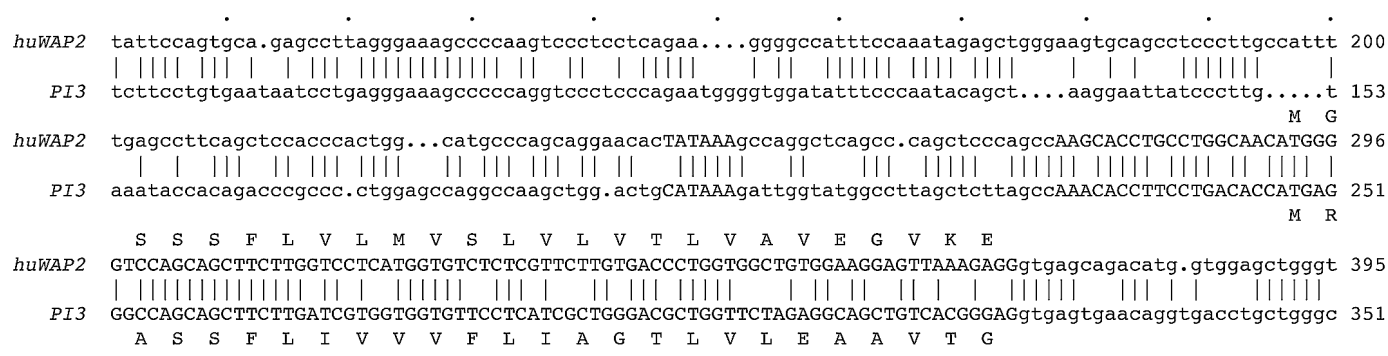

FIG. 1. A new gene with similarity to the elafin gene (*PI3*). The nucleotide sequence of the first exon and upstream promoter region of the elafin gene (*PI3*) was compared to the new gene (*huWAP2*) using the program BESTFIT in the GCG program package. TATA-box and exon sequences are written in capital letters. Translations given with the one-letter code are written above the *huWAP2* sequence and below the *PI3* sequence. Gaps are indicated by dots (.) and conserved nucleotides are indicated by vertical lines (|).

cycles at 95°C for 30 s and 68°C for 1 min. In a final step the samples were incubated at 68°C for 1 min. PCR products were analyzed by electrophoresis in 2.5% agarose gels that were stained by ethidium bromide (1 μg/ml).

RACE. Rapid amplification of cDNA ends (RACE) was done using the SMART RACE cDNA amplification kit (Clontech). First strand cDNAs for 5′ RACE and 3′ RACE were synthesized using Powerscript (Clontech) and 1 μg of prostate RNA following the protocol provided with the RACE kit. Amplification of cDNA ends was done as described in the Clontech protocol but with half of the recommended volume. Gene-specific primer for 5′ RACE was N1R1 and for 3′ RACE, N1F1. The thermal cycling protocol consisted of an initial denaturation at 95°C for 1 min, followed by 5 cycles at 95°C for 30 s and 72°C for 2 min. Another 5 cycles were run at 95°C for 30 s, 70°C for 30 s and 72°C for 2 min, followed by 30 cycles at 95°C for 30 s and 68°C for 2 min. A final step consisted of an incubation at 68°C for 2 min. RACE-products were reamplified with material obtained by dipping the tip of a micropipette in the ethidium bromide stained bands on the agarose gel. The material was transferred to a PCR tube containing 0.2 μM each of the gene-specific primer and the nested universal primer provided in the RACE kit, and additional components for PCR using Advantage 2 (Clontech) as provided with the polymerase. The PCR protocol consisted of an initial denaturation at 95°C for 1 min followed by 30 cycles at 95°C for 30 s and 68°C for 1 min, and a final extension at 68°C for 1 min. The PCR-products were then directly used as templates in DNA sequence reactions.

DNA sequencing. The DNA content in RACE-product samples were quantified by running dilutions of samples in parallel with samples of known DNA concentration on agarose gel. Approximately 100 ng RACE product served as template in DNA sequencing reactions using the Big Dye terminator cycle sequencing kit (Applied Biosystems). As sequencing primers served the above described N1F1 and N1R1 and an additional primer N1F2, AGGGTCCTGAGACTTGGAAT, that was used in order to obtain the sequence of the transcripts 3′ end. The procedure yielded a contiguous sequence without ambiguous positions, but for most part only one strand was sequenced. However, it was doomed unnecessary to sequence both strands as the generated cDNA sequence was confirmed by DNA sequences from the human genome project. Nucleotide sequences were assembled and analyzed using the GCG program package (Genetics Computer Group, Inc., WI).

RESULTS

The human semenogelin genes and the genes of the protease inhibitors elafin and SLPI are located on the overlapping genomic clones, RP1-172H20 and RP1-300I2, sequenced by the Sanger Centre. The nucleotide sequence of these clones and another clone, RP1-211D12, also overlapping with RP1-172H20, were analyzed for presence of nucleotides homologous to semenogelin and protease inhibitor genes by searching for similarities to the first exon of these genes. Approximately 52 kb upstream of *PI3*, but in opposite direction, a potentially new gene was identified with high similarity to the elafin gene (Fig. 1). As can be seen, the similarity does not only apply to the first exon as also the upstream promoter and the splice donor site of the first intron are conserved. Furthermore, the presence of a TATA-box suggests that the gene might be expressed. Around 100 bp downstream from the conserved splice donor site a potential splice acceptor site was identified and following that is a nucleotide sequence that could code for a WAP domain. To monitor whether the new gene is expressed, an RT-PCR assay was devised by which tissues samples could be screened for the presence of transcripts. As can be seen (Fig. 2) the gene is indeed transcribed as indicated by the presence of the 183-bp PCR product in several tissues. In some instances there is also a 286-bp PCR-product, probably stemming from genomic DNA that is contaminating the RNA preparations. The signal intensity relative to that of the housekeeping gene, APRT, suggests that the gene is highly expressed in prostate, skin, lung, and esophagus. Weak signals are also detected in skeletal muscle, epididymis, kidney, trachea, salivary gland, testis, and seminal vesicle. The gene is also active in the hormone sensitive prostate cell line LNCaP, but not in the hormone insensitive cell-line DU145.

To define the transcript of the new gene, RACE was conducted with cDNA made from prostate RNA. Using the oligonucleotides N1F1 and N1R1 as gene-specific primers a 5′ RACE product of around 0.3 kb and a 3′ RACE product of approximately 0.9 kb were obtained (Fig. 3). The stained band indicating a fragment 0.6 kb in the lane with the 3′ RACE product is heterogeneous

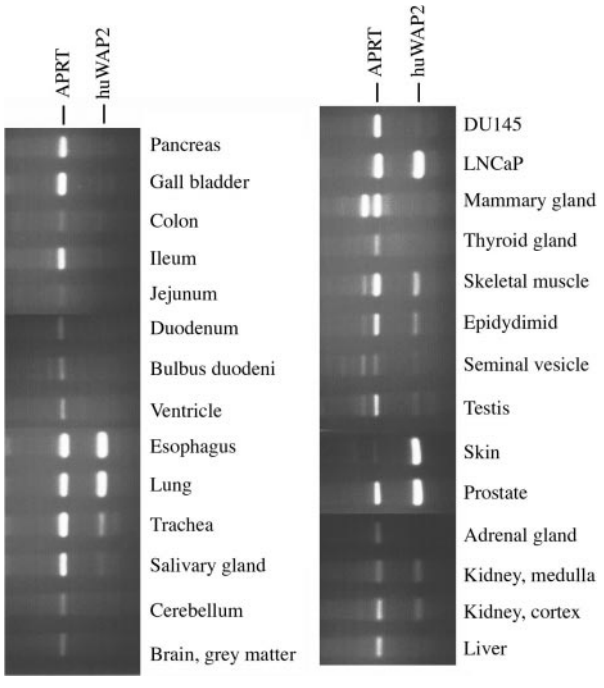

FIG. 2. Detection of transcripts by RT-PCR. RNA was extracted and cDNA synthesized from tissues as indicated. PCR was performed on cDNA using primers based on nucleotide sequences in exon 1 and exon 2. As a control, transcript-specific primers for the housekeeping gene *APRT*, encoding adenine phosphoribosyltransferase, were included in the PCR-mix. The PCR products were separated by electrophoresis in 2.5% agarose gels. Mounted photographs of ethidium bromide stained gels are shown. The upper band of 257 bp represents the ARPT transcript and the lower band of 183 bp represents the huWAP2 transcript. In a few lanes, there is also a PCR product of 286 bp that probably is derived from contaminating genomic DNA.

and probably represents an artifact. By DNA sequencing of the RACE products, the size of the new transcript was determined to 774 bp—excluding the poly(A) tail of undetermined size. By comparing the nucleotides sequence of the transcript with that of genomic DNA it can be concluded that *huWAP* contain 3 exons of 97, 159, and 518 bp, separated by introns of 103 and 163 bp (Fig. 4).

The translation of *huWAP2* yields a polypeptide of 111 amino acid residues. From the homology with the semenogelin, elafin and SLPI genes it can be predicted that residues 1 to 23 constitutes a signal peptide. The mature protein of 88 amino acid residues, tentatively called huWAP2, has a molecular weight of 9715.92 and an isoelectric point of 5.77. It is probably not glycosylated as there is no consensus sequence for N-glycosylation in the primary structure. Predominant amino acid residues are Cys and Pro (11% and 10%) as in other WAP domain-containing proteins. HuWAP2 also contains 10% Glu and 10% Lys residues and a single Trp residue that is responsible for most of the absorbance at 280 nm—the molar extinction coefficient is 8550. The secreted protein can be considered to consist of two domains, an amino-terminal WAP domain of 40 to 50 residues and a C-terminal domain of 30 to 40 residues that do not show resemblance to any of the motifs commonly present in proteins as reveled by BLAST search at the NCBI web server (http://www.ncbi.nlm.nih.gov/BLAST/).

The amino acid sequence of the WAP domain in huWAP2 was align with homologous regions of other known human WAP domain containing proteins using the program PILEUP in the GCG package (Fig. 5). As can be seen the different WAP domains are not extensively similar as in most cases only 35–50% of the amino acid residues are conserved. However, the Cys residues, which all are involved in intrachain disulfide bonds and are important for the folding of the peptide chain, are all well conserved—the exception is the second Cys residue which align poorly. The WAP domain of huWAP2 shows the highest similarity to that of elafin with 44% conserved residues.

DISCUSSION

In this report, the transcript of a new gene is described. It was initially discovered by analyses of nucleotide sequences generated by the human genome project. Judging by annotations in sequence databases, the gene has not been recognized previously neither by gene predicting computer programs nor by EST sequencing projects. The reason for this might be that the gene is relatively small and that in most tissues it is only weakly transcribed—if transcribed at all.

Several studies on elafin and SLPI show that that WAP domains can inhibit serine proteases like elastase and trypsin (11, 12). It is therefore likely that the huWAP2 protein described in this paper also functions

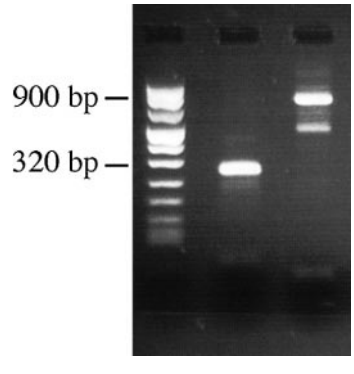

FIG. 3. Rapid amplification of cDNA ends (RACE). The ends of the huWAP2 transcript were amplified using prostate RNA and components provided by the SMART RACE cDNA amplification kit (Clontech). As gene-specific primers served the two oligonucleotides used as primers in the RT-PCR. RACE products were separated by electrophoresis in 2.5% agarose gel and visualized by staining with ethidium bromide. The lanes from left show size marker (Marker VIII, Boehringer), 5′ RACE product and 3′ RACE product. Indicated to the far left are the positions of the 900- and 320-bp size markers.

```
ctccttccataagatagtttggccctggtggcgcctaatggtcaatcacgtgtccatttcttctcctcgtgtgaagaacgagggtcgctcttggatttct  100
tcatctattccagtgcagagccttagggaaagccccaagtccctcctcagaaggggccatttccaaatagagctgggaagtgcagcctcccttgccattt  200
                                                                                    MetGlySer
tgagccttcagctccacccactggcatgcccagcaggaacactataaagccaggctcagcccagctcccagccAAGCACCTGCCTGGCAACATGGGGTCC  300
SerSerPheLeuValLeuMetValSerLeuValLeuValThrLeuValAlaValGluGlyValLysGluG
AGCAGCTTCTTGGTCCTCATGGTGTCTCTCGTTCTTGTGACCCTGGTGGCTGTGGAAGGAGTTAAAGAGGgtgagcagacatggtggagctgggtggggc  400
                                                                    lyIleGluLysAlaGlyValCysProA
tgggctggggagaggtcctgaggggccctctggggctggagttctcatatcaccttgtggcttcctccactagGTATAGAGAAAGCAGGGGTTTGCCCAG  500
laAspAsnValArgCysPheLysSerAspProProGlnCysHisThrAspGlnAspCysLeuGlyGluArgLysCysCysTyrLeuHisCysGlyPheLy
CTGACAACGTACGCTGCTTCAAGTCCGATCCTCCCCAGTGTCACACAGACCAGGACTGTCTGGGGGAAAGGAAGTGTTGTTACCTGCACTGTGGCTTCAA  600
sCysValIleProValLysGluLeuGluGluG
GTGTGTGATTCCTGTGAAGGAACTGGAAGAAGgtaaggagacctgcctcccagggctggggctgtcccttccctgcctctatctgacccatgaaagttcg  700
                                                                                    lyGly
gaggaattagtccctttagctggtgtggggagggatggctaaagctggcagggccctcagagaccacctagtctgaacatctccattgtccaaagGAGGA  800
AsnLysAspGluAspValSerArgProTyrProGluProGlyTrpGluAlaLysCysProGlySerSerSerThrArgCysProGlnLys
AACAAGGATGAAGATGTGTCAAGGCCATACCCTGAGCCAGGATGGGAGGCCAAGTGTCCAGGCTCCTCCTCTACCAGGTGTCCTCAGAAATGATGCTGGG
TCCTTTCTACCTCTGGGGGTCACTCTCACTTGGCACCTGCCCCTGAGGGTCCTGAGACTTGGAATATGGAAGAAGCAATACCCAACCCCCACCAAAGAAAA  1000
CCTGAGCTTGAAGTCCTTTTCCCCAAAAAGAGGGAAGAGTCACAAAAAGTCCAGACCCCAGGGACGGTACTTTCCCTCTCTACCTGGTGCTCCTCCCTAA
TGCTCATGAATGGACCCCTCATGAATGAAACCAGTGCCCTTATAAGAGACCCCAAAGAGCTGCCTTGCCCTTCTGCAATGTGTGATCACAGCTAGAAGGC
ACTGTCAGAGAAGAGAAACTGGTCCTCACCAGATGCTGAATCTGCTGGTGCCTTGATCTTGGACTTCCCAGCCTCTAGAACTGTAAGAAATAAATATTTG
CTGTTTATAATCCacccagtctatggtaatttgttatagcagcccaaacctgctaagacaacctaatagtaaaaaaaaaaaatcctattcaacattatca
aagtgaataaaatataaaatacctacaaataaatctagagatgtacagaaattttctgaaaggaaaccctaaaactttattaaatgacattttaaaagat  1500
```

FIG. 4. Structure of *huWAP2.* Exons were identified by comparing the nucleotide sequences of cDNA and genomic DNA. Exon sequences are given in capital letters with translations written above. The TATA-box and the polyadenylation signal are underlined.

as an inhibitor of serine proteases. On the other hand, two proteins containing WAP domains have been reported to exert other functions. The caltrin-like protein from guinea pig seminal vesicles is reported to be an inhibitor of Ca^{2+} uptake by spermatozoa and SPAI-2 isolated from porcine duodenum has been reported to be an inhibitor of Na^{+}, K^{+}-ATPase (13, 14). Also the presence of a WAP domain in the protein that is defect or missing in Kallmann's syndrome suggest that WAP domains might have other functions than being protease inhibitors (15). Because of this, biochemical studies are also required to assess the function of huWAP2. In future studies we will therefore recombinantly express the protein and study its properties as protease inhibitor.

```
elafin   KPGSCP.IILIRCAMLNPPNRCLKDTDCPGIKKCCEGSCG.MACFVPQ
SLPI-2   KPGKCP.VTYGQCLMLNPPNFCEMDGQCKRDLKCCMGMCG.KSCVSPV
huWAP2   KAGVCP.ADNVRCFKSDPPQ.CHTDQDCLGERKCCYLHCG.FKCVIPV
SLPI-1   KAGVCPPKKSAQCLRYKKPE.CQSDWQCPGKKRCCPDTCG.IKCLDPV
 HE4-2   KEGSCPQVNINFPQLGLCRDQCQVDSQCPGQMKCCRNGCGKVSCVTPN
anosmin  KQGDCPAPEKASGFAAACVESCEVDNECSGVKKCCSNGCG HTCQVPK
 HE4-1   KTGVCPELQADQN....CTQECVSDSECADNLKCCSAGCA.TFCSLPN
  ps20   RADRCPPPPRTLPPGACQAARCQADSECPRHRRCCYNGCA YACLEAV
```

FIG. 5. Alignment of human WAP domains. The amino acid sequences of WAP domains were aligned using the program PILEUP in the GCG program package. The compared sequences are from the single WAP domain-containing proteins elafin, huWAP2, anosmin—the protein that is missing in patients with Kallmann's syndrome, ps20—prostate stromal protein of 20 kDa, and the two WAP domains in secretory leukocyte protease inhibitor (SLPI-1 and SLPI-2) and human epididymis gene product HE4 (HE4-1 and HE4-2).

ACKNOWLEDGMENTS

The assistance of I. Dahlquist and S. Strömberg in running the automated DNA sequencer is acknowledged. This work was supported by a grant from the Swedish Medical Research Council (Project 8660).

REFERENCES

1. Lundwall, Å., and Lazure, C. (1995) A novel gene family encoding proteins with highly differing structure because of a rapidly evolving exon. *FEBS Lett.* **374,** 53–56.
2. Lilja, H., Abrahamsson, P. A., and Lundwall, A. (1989) Semenogelin, the predominant protein in human semen. Primary structure and identification of closely related proteins in the male accessory sex glands and on the spermatozoa. *J. Biol. Chem.* **264,** 1894–1900.

3. Lilja, H., and Lundwall, A. (1992) Molecular cloning of epididymal and seminal vesicular transcripts encoding a semenogelin-related protein. *Proc. Natl. Acad. Sci. USA* **89,** 4559–4563.
4. Ulvsbäck, M., Lazure, C., Lilja, H., Spurr, N. K., Rao, V. V. N. G., Löffer, C., Hansmann, I., and Å., L. (1992) Gene structure of semenogelin I and II. The predominant proteins in human semen are encoded by two homologous genes on chromosome 20. *J. Biol. Chem.* **267,** 18080–18084.
5. Lundwall, Å., and Ulvsbäck, M. (1996) The gene of the protease inhibitor SKALP/elafin is a member of the REST gene family. *Biochem. Biophys. Res. Commun.* **221,** 323–327.
6. Saheki, T., Ito, F., Hagiwara, H., Saito, Y., Kuroki, J., Tachibana, S., and Hirose, S. (1992) Primary structure of the human elafin precursor preproelafin deduced from the nucleotide sequence of its gene and the presence of unique repetitive sequences in the prosegment. *Biochem. Biophys. Res. Commun.* **185,** 240–245.
7. Stetler, G., Brewer, M. T., and Thompson, R. C. (1986) Isolation and sequence of a human gene encoding a potent inhibitor of leukocyte proteases. *Nucleic Acids Res.* **14,** 7883–7896.
8. Hennighausen, L. G., and Sippel, A. E. (1982) Mouse whey acidic protein is a novel member of the family of 'four-disulfide core' proteins. *Nucleic Acids Res.* **10,** 2677–2684.
9. Molhuizen, H. O. F., Zeeuwen, P. L. J. M., Olde Weghuis, D., Geurts van Kessel, A., and Schalkwijk, J. (1994) Assignment of the human gene encoding the epidermal serine proteinase inhibitor SKALP (PI3) to chromosome region 20q12-q13. *Cytogenet. Cell Genet.* **66,** 129–131.
10. Chomczynski, P., and Sacchi, N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. *Anal. Biochem.* **162,** 156–159.
11. Wiedow, O., Schroder, J. M., Gregory, H., Young, J. A., and Christophers, E. (1990) Elafin: An elastase-specific inhibitor of human skin. Purification, characterization, and complete amino acid sequence. *J. Biol. Chem.* **265,** 14791–14795.
12. Thompson, R. C., and Ohlsson, K. (1986) Isolation, properties, and complete amino acid sequence of human secretory leukocyte protease inhibitor, a potent inhibitor of leukocyte elastase. *Proc. Natl. Acad. Sci. USA* **83,** 6692–6696.
13. Coronel, C. E., San Agustin, J., and Lardy, H. A. (1990) Purification and structure of caltrin-like proteins from seminal vesicle of the guinea pig. *J. Biol. Chem.* **265,** 6854–6859.
14. Araki, K., Kuroki, J., Ito, O., Kuwada, M., and Tachibana, S. (1989) Novel peptide inhibitor (SPAI) of Na^+, K^+-ATPase from porcine intestine. *Biochem. Biophys. Res. Commun.* **164,** 496–502.
15. Legouis, R., *et al.* (1991) The candidate gene for the X-linked Kallmann syndrome encodes a protein related to adhesion molecules. *Cell* **67,** 423–435.