LUND UNIVERSITY

A locus on human chromosome 20 contains several genes expressing protease inhibitor domains with homology to whey acidic protein.

Clauss, Adam; Lilja, Hans; Lundwall, Åke

Published in:
Biochemical Journal

DOI:
10.1042/BJ20020869

2002

Link to publication

Citation for published version (APA):
Clauss, A., Lilja, H., & Lundwall, Å. (2002). A locus on human chromosome 20 contains several genes expressing protease inhibitor domains with homology to whey acidic protein. *Biochemical Journal*, *368*(Pt 1), 233-242. https://doi.org/10.1042/BJ20020869

Total number of authors:
3

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

A locus on human chromosome 20 contains several genes expressing protease inhibitor domains with homology to whey acidic protein

Adam CLAUSS, Hans LILJA and Åke LUNDWALL[1]

Wallenberg Laboratory, Department of Laboratory Medicine, University Hospital MAS, Lund University, S-205 02 Malmö, Sweden

A locus containing 14 genes, encoding protein domains that have homology with whey acidic protein (WAP), has been identified in a region of 678 kb on human chromosome 20q12-13.1. Among them are genes of the known or postulated protease inhibitors elafin, secretory leucocyte protease inhibitor, human epididymis gene product 4, eppin, and huWAP2. Nucleotide sequences of full-length transcripts were obtained from cDNA fragments generated by rapid amplification of cDNA ends. Characteristic features of the genes are that the upstream promoter regions are devoid of TATA-boxes and that the coding nucleotides are divided into distinct exons for the signal peptide and for each WAP domain. In most cases, there is also a separate exon encompassing a few terminal codons and the 3′ untranslated nucleotides. There are also examples of mixed type inhibitors, that encode inhibitor domains of both WAP and Kunitz types. Several of the genes appear to be expressed ubiquitously, but, in most cases, the highest transcript levels are found in epididymis followed by testis and trachea. Some of the genes also display high transcript levels in neural tissues. Potential biological roles of protein products could be in host defence against invading micro-organisms or in the regulation of endogenous proteolytic enzymes, of which those originating from the kallikrein gene locus on chromosome 19 are of particular interest.

Key words: epididymis, serine protease, transcript.

INTRODUCTION

An immense number of biological activities are controlled by peptide bond cleavages effected by proteolytic enzymes. Examples range from food digestion to the highly specific activation of biochemical cascades. Serine proteases (EC 3.4.21) form a diverse family of proteolytic enzymes that carry conserved histidine, aspartate and serine residues at the catalytic site. Well-known examples are components of the blood coagulation cascade, the complement system, and the digestive enzymes trypsin and chymotrypsin. These enzymes are typically synthesized and secreted as latent, inactive zymogens that, after limited proteolytic modification, manifest their proteolytic action in appropriate settings, whereas careful control of inappropriate exposure of this activity is governed by tight regulation by protease inhibitors.

There are several different types of serine protease inhibitors that vary in specificity and inhibitory capacity. Serpins, like α-1-protease inhibitor and antithrombin III, are efficient inhibitors that, following proteolytic attack, covalently bind to the target serine protease and destroy the structure at the catalytic site [1]. Other inhibitors, such as the small Kunitz and Kazal inhibitor types, bind to the catalytic cleft of the target enzyme, thereby preventing access by substrates [2,3]. The same mode of inhibitory mechanism is exerted by the four-disulphide core inhibitor type. These small inhibitors consist of domains of approx. 50 amino acid residues in size that, as the name implies, have a characteristic disulphide pattern. Because the motif was first identified in whey acidic protein (WAP), a predominant protein in milk whey of lactating mice, they are also known as WAP domains [4]. The most-studied of the human WAPs are secretory leucocyte protease inhibitor (SLPI) and elafin which are both potent inhibitors of leucocyte elastase [5,6]. Three-dimensional structures are known for both elafin, in complex with porcine pancreatic elastase, and SLPI, in complex with α-chymotrypsin [7,8]. The flat and slightly twisted WAP domain consists of a central core β-sheet surrounded by two external peptide chains that are connected by a loop that holds the protease binding site.

In the literature, there are also reports that suggest that WAP domains might not only be protease inhibitors. Porcine SPAI-2 (sodium–potassium ATPase inhibitor-2) has, for instance, been claimed to inhibit an intestinal Na^+/K^+ ATPase and a caltrin-like protein secreted by the guinea pig seminal vesicles is reported to inhibit Ca^{2+} uptake by spermatozoa [9,10]. The protease inhibitor SLPI has also been demonstrated to be a potent anti-microbial agent, a function that is claimed to be separated from the anti-protease activity, as reviewed in [11]. Another WAP domain of unknown function is present in the large protein that is defective or missing in Kallmann syndrome [12].

It has previously been shown that elafin and SLPI are related to the major gel-forming proteins of human semen, semenogelin I and semenogelin II [13]. Although the semenogelins and the WAPs are structurally very different, their genes are similarly organized and conserved nucleotide sequences are present in the first and the last exon. In contrast, the central exons, which carry

Abbreviations used: APRT, adenine phosphoribosyltransferase; EST, expressed sequence tag; HE4, human epididymis gene product 4; PSA, prostate-specific antigen; RACE, rapid amplification of cDNA ends; RT-PCR, reverse transcriptase-PCR; SLPI, secretory leucocyte protease inhibitor; WAP, whey acidic protein.

[1] To whom correspondence should be addressed (e-mail Ake.Lundwall@klkemi.mas.lu.se).

The nucleotide sequences reported in this paper have been submitted to the GenBank®, EMBL, DDBJ and GSDB Nucleotide Sequence databases under accession numbers AY038181 for WAP1, AF411861 for WAP6, AF492015 for WAP8a, AF492016 for WAP8b, AY047610 for WAP9, AY038182 for WAP10, AY047609 for WAP11, AF454506 for WAP12a, AF454507 for WAP12b, AF454505 for WAP13 and AF488306 for WAP14.

© 2002 Biochemical Society

almost all of the coding nucleotides except those for the signal peptide, are greatly different, thereby explaining the lack of structural conservation of the secreted proteins [14]. The genes of semenogelin I (*SEMG1*), semenogelin II (*SEMG2*) and elafin (*PI3*) have previously been localized to chromosome 20q12-13.1 [15,16], and, with the advent of nucleotide sequences from the Human Genome Project, it has been possible to assign *PI3*, *SEMG1*, *SEMG2* and *SLPI* to a single locus, approx. 80 kb in size. By searching in the vicinity of this locus for nucleotide sequences with similarity to those of the first exon of *PI3*, *SEMG1*, *SEMG2* and *SLPI*, a novel gene encoding a WAP domain, *WAP2*, was identified 52 kb upstream of *PI3* [17]. In the present paper, we extend our search for new genes encoding WAP domains by surveying a larger region on chromosome 20.

EXPERIMENTAL

Identification of candidate genes

Starting with clone RP1-172H20, containing *PI3*, *SEMG1* and *SEMG2*, overlapping DNA sequences in GenBank® were identified by BLAST searches using 100 bp of clone terminal sequence as the query. Sequences were assembled into contigs of around 700 kb and translated in six reading frames using the freeware computer program DNAid (downloaded from http://homepage.mac.com/cellbiol/soft.htm). Using the program's search function, the translations were surveyed for the motifs Cys-Xaa-Xaa-Xaa-Xaa-Xaa-Cys-Cys and Cys-Cys-Xaa-Xaa-Xaa-Cys, which are part of the conserved disulphide pattern. Sequences that were identified by the partial motif search were then inspected manually for similarity to the full cysteine motif of WAP domains as defined in the Prosite database (http://www.expasy.ch/prosite/).

Table 1 PCR primers

Nucleotide sequences of oligonucleotides used as primers for RT-PCR. Primer pairs are notated with the forward primer above the reverse primer. The transcript length indicates the size of the PCR product generated by priming on a spliced transcript.

Gene	Primer sequence	Transcript length (bp)
WAP1	5′-TGACCAGTGCGTGGAAGACAGCCAGTGT-3′	201
	5′-CAGGCGCTGTGGCAGCATCGCTTTTTGC-3′	
PI3	5′-CTGCTTGAAAGATACTGACTGCCCAGGAAT-3′	199
	5′-ATCCTGAATGGGAGGAAGAATGGACAGTGT-3′	
SLPI	5′-ACTGGCAGTGTCCAGGGAAGAAGAGATGTT-3′	215
	5′-GGATTTCCCACACATGCCCATGCAACACTT-3′	
HE4	5′-GACAGCGAATGCGCCGACAACCTCAAGT-3′	209
	5′-GTGACACAGGACACCTTCCCACAGCCAT-3′	
WAP6	5′-ATGGGACTCTCAGGACTTCTGCCAATCC-3′	198
	5′-TCCACGGCTGAACGGGCAACACTTCATGTT-3′	
Eppin	5′-ACAAGAAGTGTTGTGTCTTCAGCTGCGG-3′	163
	5′-TGGCAGCCACCATAGACAAACATGGAGC-3′	
WAP8	5′-CCACCTTCATGTCACAGTGACATCGATT-3′	175
	5′-ATGGGCACTCCTCATCCTGCAGGCACTT-3′	
WAP9	5′-TGCTGGGTACAGCCTCCATATAAGTACT-3′	212
	5′-CAGCACTCTTCTTAGACCAGATGGTCAT-3′	
WAP10	5′-AGGCCCAGGGAGGATACCGTGACAAGAA-3′	200
	5′-CCACTCTCCCACTCACAGGATGCTCATG-3′	
WAP11	5′-TACGGTGCTACTGTCTGTGCTGGGAGAAAT-3′	181
	5′-GCAGATGTTTCCACAATAGGTCCAGCAGCAT-3′	
WAP12	5′-GTGACAAGATGAGGATGCAGAGAATCAA-3′	145
	5′-ATGCTCATACAAATGTTCCCACAGAAGG-3′	
WAP13	5′-AGTATATCTTGGAACCTCCACCCTGCATAT-3′	196
	5′-CCTCAGTTGGCAGGCATGATGACTTCT-3′	
WAP14	5′-GCTGTGTGATGGATGAGAATTGTCAAGCT-3′	181
	5′-AAGAAGCTCCAGACAAATCAGCACAGCTA-3′	
APRT	5′-GCCGCATCGACTACATCGCAGGCCTAGA-3′	257
	5′-CTCACAGGCAGCGTTCATGGTTCCACCA-3′	

Detection of transcripts

The tissue distribution of transcript was investigated using reverse transcriptase-PCR (RT-PCR). Preparation of RNA samples and conditions for cDNA synthesis are described in [17]. Approx. 3 μg of total RNA served as template for cDNA synthesis in a volume of 15 μl, of which 1 μl was subsequently used for PCR in a total reaction volume of 10 μl. Enzyme and buffer components for PCR were provided by the Advantage 2 kit (Clontech, Palo Alto, CA, U.S.A.). Each PCR reaction was run with 0.2 μM primers for a WAP transcript and for adenine phosphoribosyltransferase (APRT), which served as an internal control. Primers were selected from nucleotide sequences that were located on separate exons of their respective genes, so that the size of PCR products differed between the spliced transcripts and chromosomal DNA or unspliced transcripts (Table 1). PCR was performed using a PTC-200 Peltier Thermal Cycler (MJ Research, Watertown, MA, U.S.A.) with a program consisting of an initial denaturation at 95 °C for 1 min, followed by 35 cycles at 95 °C for 30 s (denaturation) and 68 °C for 1 min (annealing and extension). Finally, the samples were incubated at 68 °C for 1 min. The PCR products were analysed by electrophoresis in 2.5 % agarose gels containing 1 μg/ml ethidium bromide. The gels were inspected and photographed under 305 nm UV light.

Determination of transcript structures

Transcripts of the novel genes were characterized by the sequencing of products generated by rapid amplification of cDNA ends (RACE). As templates to cDNA synthesis, RNA from oesophagus and testis was used for WAP1 and WAP14, while epididymis RNA provided a source for all other transcripts. The RACE procedure was performed essentially as described in [17], using the SMART RACE cDNA amplification kit (Clontech, Palo Alto, CA, U.S.A.). The primers were, in most cases, the same as used for RT-PCR, which yielded overlapping 5′ RACE and 3′ RACE products from which the complete nucleotide sequences of transcripts were deduced. Products of the RACE reactions were separated by electrophoresis using 2.5 % agarose gels and ethidium-bromide-stained bands were recovered and purified on Jetsorb resins (Genomed, Bad Oeynhausen, Germany). The DNA concentration of each RACE product was estimated from the staining intensity, relative to standard samples, following electrophoresis in agarose gels containing ethidium bromide. DNA sequencing was performed directly on PCR products using the Big Dye Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems, Foster City, CA, U.S.A.) and approx. 100 ng of template DNA. The RT-PCR primers or specific 20-mer oligonucleotides designed from known DNA-sequences were used for sequencing primers. Nucleotide sequences were assembled and analysed using the GCG computer program package (Genetics Computer Group, Madison, WI, U.S.A.). Signal peptidase cleavage sites were predicted by the computer program SignalP from the Center for Biological Sequence Analysis website (http://www.cbs.dtu.dk/services/SignalP/).

© 2002 Biochemical Society

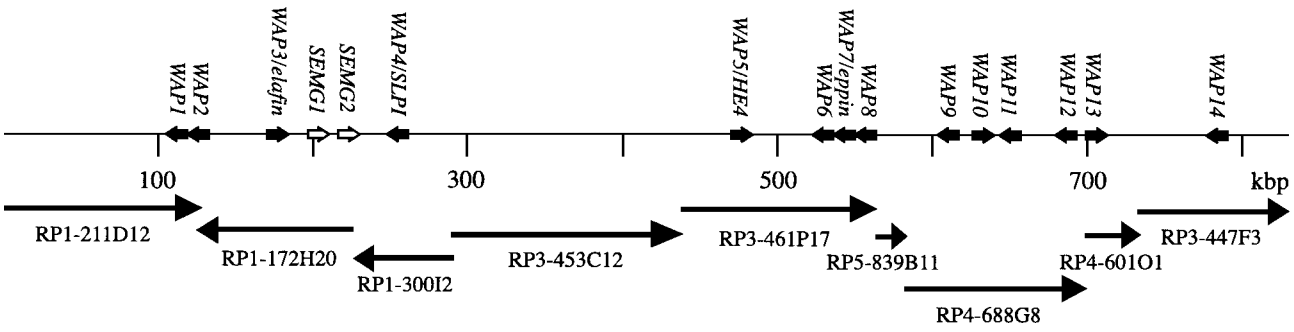

Figure 1 Schematic drawing of the WAP-locus

The upper line illustrates genomic DNA, overlapping with the cytogenic chromosome region 20q12-13, on which equally sized arrows indicate the approximate location of genes. The arrows in the lower part of the figure indicate the location of clones that were sequenced as a part of the Human Genome Project and which were used to assemble a contiguous DNA sequence.

RESULTS

A locus with genes encoding WAP domains

Overlapping DNA sequences from clone RP5-881L22 to clone RP11-465L10 yielded a contiguous sequence of 1.8 Mb, overlapping with the cytogenic region q12-13.1 on chromosome 20. The 100 bp overlap between clones RP4-601O1 and RP3-447F3 was ambiguous because of two mismatches and a stretch of 6 bp that is present only in RP4-601O1. The difference is probably the result of a polymorphism, as PCR of human genomic DNA, using primers based on the sequence at the clone's end, yielded a single product of a size indicating that the clones overlapped by 100 bp. This was also confirmed by the nucleotide sequence of the PCR product, which was identical to that of RP3-447 over the differing region.

The assembled DNA sequence was translated in six reading frames that were surveyed for WAP motifs. In total, 22 motifs were found in a 678 kb region that started 65 kb on the centromeric side of *PI3* and extended in the telomeric direction. By sequence analyses and RT-PCR, the motifs could be assigned to 14 genes denoted *WAP1* to *WAP14* (Figure 1). Of these, hereafter called WAP genes, *WAP2* was reported recently and *WAP3*, *WAP4*, *WAP5*, and *WAP7* encode the known proteins elafin, SLPI, human epididymis gene product 4 (HE4), and eppin respectively [5,6,17–19]. The remaining nine genes have not previously been described in the literature.

Organization of the WAP locus

The WAP genes are clustered in two sub-loci that are separated by a region of 215 kb, containing unrelated genes, for example, *SDC4* and *MATN4*, which encode the heparan sulphate proteoglycan, syndecan 4, and the extracellular matrix protein, matrilin 4. The centromeric cluster of four WAP genes is distributed along a region of 145 kb, which also holds the semenogelin genes (Figure 1). Of the WAP genes in the cluster, only *WAP1* has not been described previously. The overall organization of this gene is very similar to that of *SLPI*, with two exons encoding WAP domains and only two protein-coding nucleotides located in the last exon. However, the resemblance does not extend to the nucleotide sequences, as the sequence similarity is less than 40 %.

The telomeric cluster contains the two epididymis-transcribed genes, *HE4* and *Eppin*, and eight new WAP genes (Figure 1). Two of the new genes, *WAP6* and *WAP8*, are located close to *Eppin*, to which they also are more structurally related than to the other WAP genes. *Eppin* and *WAP6* were probably formed by a gene duplication, because the nucleotide sequence of a region extending from around 0.5 kb upstream of the translation initiation codon to around 0.5 kb downstream of the stop codon is conserved by 60–65 %. In the case of *Eppin* and *WAP8*, there is no extensive sequence similarity but both genes code for a Kunitz-inhibitor domain in addition to the WAP domains.

The region encompassing *WAP9–WAP13* constitutes a block of genes that encode single WAP-domains. A comparison of nucleotide sequences shows that, generally, the genes are poorly conserved. However, in the case of *WAP10* and *WAP12*, a 3-kb region, extending approx. 0.4 kb upstream of the translation initiation codons to around 1.4 kb downstream of the poly-adenylation site, is 87.5 % conserved, indicating that they were formed by a relatively recent gene duplication. In addition, *WAP9* and *WAP11* are likely to have been formed by a gene duplication that can be traced through conserved nucleotide sequences. In this case, 2 kb regions that start approximately at the transcription initiation site and end a few nucleotides before the poly-adenylation signal display a sequence similarity of 63 %, suggesting that this duplication occurred earlier than the one that gave rise to *WAP10* and *WAP12*.

Structure of transcripts

The products of 5′ RACE did not give rise to nucleotide sequences with distinct 5′ ends; instead, the sequences gradually faded away, which suggests imprecise transcription initiation sites. Probably, this is because none of the newly discovered genes carry a TATA-box in the proximal promoter. The nucleotide sequences of transcripts with translated amino acid residues are shown in Figure 2. The sequence lengths roughly agree with transcript sizes as calculated from the mobility of RACE products upon electrophoresis in agarose gels. Heterogeneity arising from alternative splicing was observed in transcripts from *WAP8*, *WAP12* and *WAP14*. The two *WAP8* transcripts differ only in 3′ untranslated nucleotides. The difference between the two *WAP12* transcripts is more spectacular, as nucleotides derived from the exon that encodes the signal peptide in the larger transcript (*WAP12a*) are missing in the smaller transcript (*WAP12b*). Thus the protein that is synthesized from the smaller transcript may not be secreted. The transcription of *WAP14* is very complex and generates several different transcripts that will be reported in

© 2002 Biochemical Society

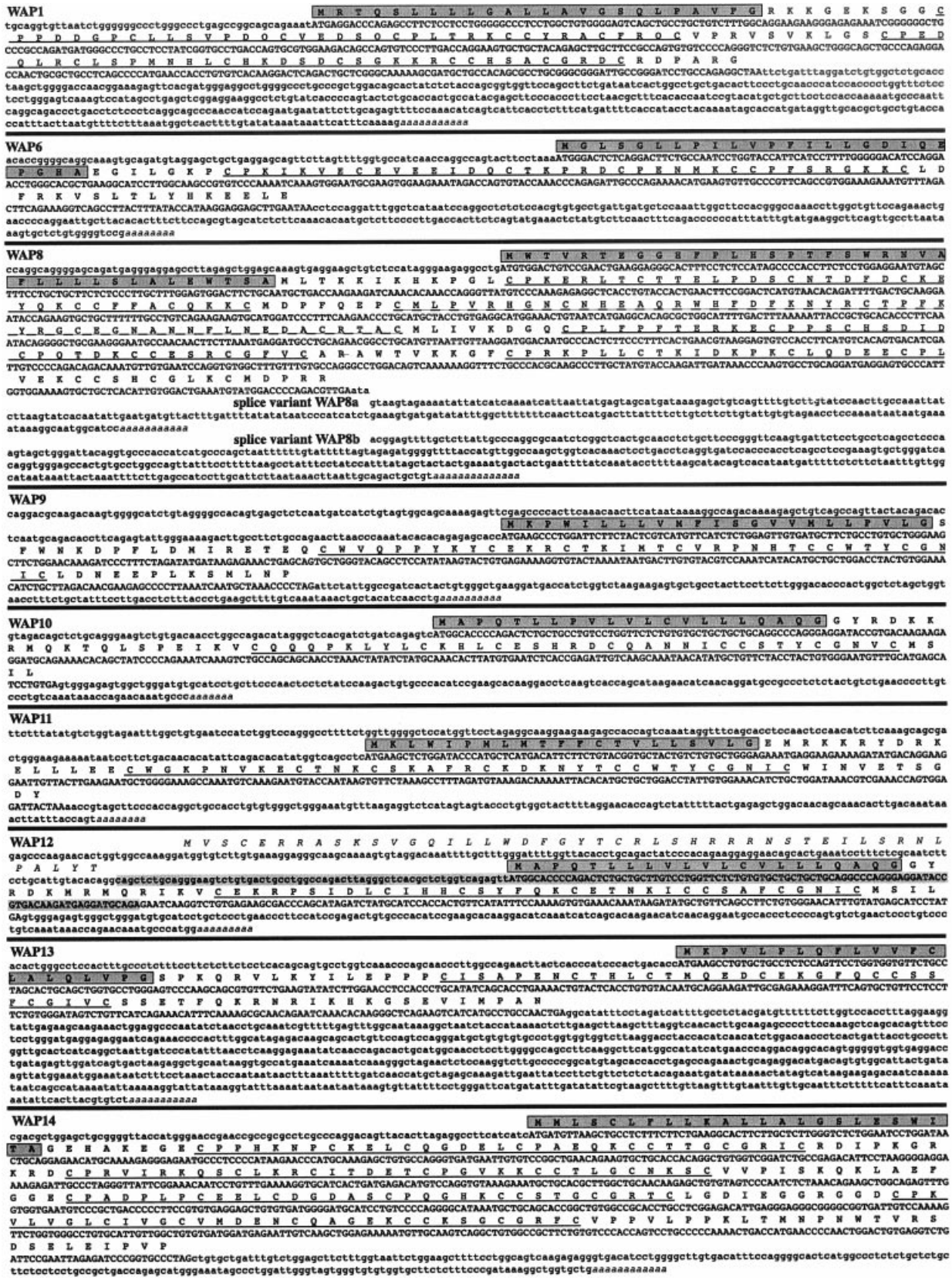

Figure 2 Transcript sequences

The nucleotide sequences of WAP gene transcripts are shown with translations in one-letter code written above. Proposed signal peptides are indicated by shaded boxes. The WAP motifs are underlined with solid lines and the Kunitz domain of *WAP8* is indicated by a broken line. The nucleotides that are specific for the large transcript of *WAP12* (WAP12a) are shaded and the N-terminal translation that is specific for the short transcript (WAP12b) is indicated with italics. The first residue of the common translation differs between transcripts – the arginine of WAP12a shown in the figure is replaced by a glutamine in WAP12b.

© 2002 Biochemical Society

Table 2 ESTs

Sequences in the EST database at NCBI with great similarity to transcripts of WAP genes are listed.

Gene	Transcript size (bp)	GenBank® EST	Position in transcript	Similarity (%)	Source
WAP1	963	BG675206	1–900	97	Skin
		AF147078	454–962	99	Colon cancer
		AI307134	591–963	99	Mixed
		AI242082	592–963	99	Mixed
		BF353401	536–723	98	Mixed
		BF349114	596–752	98	Mixed
WAP6	619	AL449631	180–619	99	Testis
		BG216922	180–619	99	HT1080 cells
		BF376055	235–550	100	Testis
		BF376068	246–550	99	Testis
		AW183270	393–603	97	Mixed
		BE044426	180–312	100	Mixed
		BI519942	180–311	99	Brain
*WAP8** common sequence	803	AL449583	25–299	98	Testis
		AL449584	50–315	99	Testis
		AL449587	108–497	99	Testis
		AL449581	264–768	99	Testis
		AL449578	246–766	99	Testis
		AL449579	244–791	99	Testis
Short (a) transcript	1063	AA626150	653–1063		Testis
		AA885204	748–1063		Mixed
		AL449565	809–1063		Testis
Long (b) transcript	1263	AL449561	783–1263		Testis
		AL449564	814–1263		Testis
		AL449562	806–1259		Testis
WAP9	669	AW183195	121–669	100	Mixed
		AW269752	183–669	100	Mixed
		AW469819	283–669	98	Mixed
WAP10	481	BG195907	1–467	99	HT1080 cells
		BG195906	1–466	99	HT1080 cells
		BG194425	1–453	99	HT1080 cells
		AI673291	77–481	99	Uterus
		BG181924	1–392	95	HT1080 cells
WAP11	614	No EST sequences			
WAP12	634	AI808149	425–634	98	Mixed
		AI808159	425–634	97	Mixed
		BI035811	452–634	96	Nervous tumour
WAP13	1369	No EST sequences			
WAP14	993	AL449630	112–676	99	Testis
		AI985094	604–993	99	Kidney
		BF197609	Alternatively spliced		Kidney
		AW448916	619–993	99	Colon
		AI243277	626–993	99	Mixed
		BI460167	Alternatively spliced		Testis
		AL449629	112–345	98	Testis
		AA536130	Alternatively spliced		Prostate
		BI912742	Alternatively spliced		Brain

* In total, 28 ESTs of *WAP8* were found: 11 for *WAP8a* transcripts, five for *WAP8b* transcripts and 12 with sequences confined to the part that is common to the two *WAP8* transcripts.

more detail elsewhere. The *WAP14* transcript shown in Figure 2 corresponds to that carrying the complete set of four WAP domains.

The subset of GenBank® containing expressed sequence tags (ESTs) was probed by the nucleotide sequences of the novel transcripts using a standard BLAST search. In this way, several ESTs were found that confirmed the transcripts identified in the present study (Table 2). Although there are some minor sequence differences, the ESTs confirm the full sequences of transcripts from *WAP1*, *WAP10* and both *WAP8* transcripts. In the case of *WAP9*, the ESTs cover all of the translated nucleotides, but only part of the 5′ untranslated nucleotides. As described below, *WAP9* carries two exons encompassing 5′ untranslated nucleotides and most of the sequence in the second of these exons is also covered by ESTs. In *WAP12*, the ESTs merely cover the 3′ untranslated nucleotides and some 30 translated nucleotides, whereas there are no ESTs at all in the database from *WAP11* or *WAP13*. An interesting set of ESTs was retrieved by searching with the *WAP6* transcript. None of them cover the first exon of the gene; instead, there are three ESTs indicating that exons encoding the Kunitz domain in *Eppin* and the WAP domain in *WAP6* are joined by splicing. This very surprising finding was investigated further by RT-PCR. Using primers derived from *Eppin* and *WAP6*, the presence of a mixed transcript was

© 2002 Biochemical Society

Table 3 WAP gene splicing

Sizes of exons and introns, together with nucleotide sequences at the splice donor and acceptor sites are shown. Exon sequences are shown in capital letters, while intron sequences are shown in lower case letters.

Gene	Exon	Exon size (bp)	Splice acceptor	Splice donor	Intron size (bp)
WAP1	1	134	GAAGGGAGgtgagt		4223
	2	131	ttacagAGAAATCG	GGTCTCTGgtaaat	94
	3	141	ttgcagTGAAGCTG	TGCCAGAGgtacgg	386
	4	547	cctcagGCTAATTC		
WAP6	1	179	CCTTGGCAgtaagt		1222
	2	131	ccttagAGCCGTGT	TCAGAAAGgtaact	3458
	3	309	tcacagGTCAGCCT		
WAP8	1	105	GAAGGAGGgtgagg		17002
	2	110	atacagGCACTTTC	GATCAAACgtgagt	3117
	3	141	ttgcagACAAACCA	CTTTCAAGgtagcg	2983
	4	168	cctcagAACCCTGC	GTTAATTGgtgaga	2424
	5	141	ttttagTTAAGGAT	CTGGACAGgtaagg	970
	6a	403	ataaagTCAAAAAA		
	6b	143	ataaagTCAAAAAA	GTTGAATAgtaagt	410
	7	460	tgtgagACGGAGTT		
WAP9	1	77	CAAAAGAGgtagga		16481
	2	94	ccccagTTCGAGCC	CTTCAGAGgtaggt	4388
	3	149	tagcagTATTGGGA	AGATCCCTgtgggt	1280
	4	148	tctcagTTCTAGAT	GACAACGAgtgagt	522
	5	202	ttctagAGAGCCCC		
WAP10	1	159	GATGCAGAgtaggt		965
	2	322	tgtcagAAACACAG		
WAP11	1	62	TCTGGTTGgtaagt		3008
	2	82	gtgcagGGGCTCCA	CTTCAAAGgtaggt	16436
	3	151	ctctagCAGCGACT	ATATGACAgtaggt	1101
	4	143	ttttcagGGAAGGAA	TAAACGTCgtgagt	517
	5	176	ttacagGAAACCAG		
WAP12	1	101	ACCTGCAGgtacag		356
	2*	65	gaacagACTATCCC	GTACACAGgtgagt	18440
	3	155	ccacagGCAGCTCT	GATGCAGAgtaggt	941
	4	309	ttttcagGAATCAAG		
WAP13	1	195	TGTTCTGAgtaggt		2232
	2	151	ttccagAGTATATC	CGCAACAGgtaaga	1268
	3	65	ctacagAATCAAAC	ATCATTTTgtaagt	1930
	4	958	ccacagGCCTCTAC		
WAP14	1	76	GAGGCCTTgtaggt		1849
	2	89	tctcagCATCATCA	AGAACATGgtgagc	834
	3	129	ttgcagCAAAAGAG	TCCTAAGGgtatgt	948
	4	147	ctacagGGAGGAAA	GAAGCTGGgtaaga	10726
	5	135	atgtagCAGAGTTT	TGAGGGAGgtatgt	1472
	6	186	ccacagGGCGGGGC	CGAATTAGgtgagt	972
	7	231	ttgcagAGATCCCG		

* Exon skipped in transcript WAP12b.

confirmed. It showed the same tissue distribution as the normal *Eppin* transcript as described below.

WAP genes

Nucleotide sequences at splice donor and acceptor sites along with exon and intron sizes are shown in Table 3. Exons and introns were localized by comparing nucleotide sequences of transcripts to that of genomic DNA. Common features are that the WAP domains are coded for by single exons, which are all translated in the second reading frame (phase 1 exon). Except for *WAP8*, all WAP genes carry sequences that encode the signal peptide and the N-terminus of the mature protein on a single exon. Most of the genes also carry a last exon with only a few translated nucleotides followed by stop codon and a couple of hundred 3′ untranslated nucleotides. In the case of *WAP10*, *WAP12* and *WAP8a*, there is no separate exon with 3′ untranslated nucleotides, as these follow immediately after the nucleotides coding for the WAP domain. Also, *WAP13* differs slightly from the majority of genes at the locus, as it carries a small coding exon between the exon encoding the WAP domain and the exon carrying 3′ untranslated nucleotides.

Dissimilar to other WAP genes within the locus, *WAP9*, *WAP11* and *WAP12* carry two exons that encompass 5′ untranslated nucleotides. As these exons are separated from the coding exons by fairly large introns, the sizes of the genes by far exceed those of most of the other genes at the locus. Interestingly, none of the duplications at the locus appears to include the 5′ untranslated exons and thus they were probably recruited to the genes at a later stage. One effect of the large introns is that the genes have become intertwined, e.g. the entire *WAP10* is located within the first intron of *WAP9* (Figure 3A). As they are encoded by opposite strands, the poly-adenylation site of *WAP10* is located only 11 bp downstream from the 3′ splice site of the first exon in *WAP9*. Even more complex is the interaction between *WAP12* and *WAP13*; these genes are also transcribed in opposite directions, and, moreover, they also share exon sequences (Figure 3). Part of the WAP-encoding exon of *WAP13* is also used in *WAP12* to generate 5′ untranslated nucleotides in the larger transcript of the gene. In the small transcript of *WAP12*, these nucleotides might also be translated, leading to the unusual situation for chromosomal genes that both DNA strands are translated at this location.

Characteristics of WAP-type proteins

All of the novel transcripts reported in this paper are relatively small, ranging in size from 481 to 1369 bp, and thus they are also translated into relatively small proteins, ranging in size from 73 to 231 amino acid residues for the precursor molecules. The signal peptidase cleavage sites were computed and then evaluated by comparison with the experimentally verified sites in semenogelin I, semenogelin II, elafin, and SLPI. WAP8 was excluded from the comparison as its gene carries the codons of the signal peptide on two exons. The comparison shows that the proposed signal peptidase cleavage sites are located at homologous positions, except for WAP9. The most likely position for the N-terminus of mature WAP9 is six residues downstream of the site that was proposed by the computer program. The proposed signal peptides are indicated in Figure 2. Additional predicted features of the mature proteins are shown in Table 4.

The amino acid sequences of all WAP domains at the locus on chromosome 20 were aligned and analysed with regard to the presence of conserved residues. Except for the cysteine residues, the primary structures are not very well conserved (Figure 4). In particular, the sequence that encompasses residues 11–26, which, in elafin, contains the protease-binding site, is highly variable, suggesting that the proteins are active against a broad spectrum of proteolytic enzymes. Residues 40–54, which, in elafin, fold to the β-sheet at the centre of the molecule, constitute the most conserved part of the motif, as expected. However, the rather high frequency of substitutions, even at the core of the proteins, suggest that they were formed relatively early during vertebrate evolution. It is also interesting to note that six of the 22 aligned sequences (i.e. 27%) do not comply with the consensus pattern for WAP domains in the Prosite database. WAP6 differs from the other proteins in that the seventh cysteine residue of the conserved cysteine motif is mutated to a serine residue. In elafin, the analogous residue is located in the hairpin loop that connects the two strands of the β-sheet at the core of the

© 2002 Biochemical Society

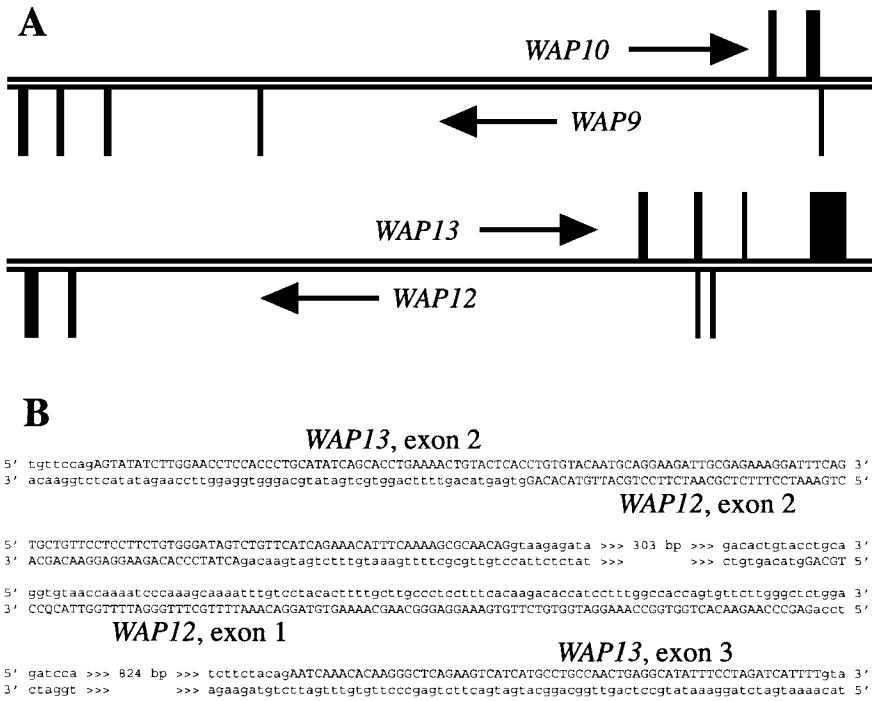

Figure 3 Overlapping genes

(**A**) Schematic drawing that shows overlapping genes with location of exons indicated by filled boxes. Arrows show direction of transcription. (**B**) Nucleotide sequence covering the overlapping exon 2 in *WAP12* and *WAP13*. Exon sequences are denoted using capital letters and intron sequences are indicated by lower case letters.

Table 4 Molecular properties of predicted WAP proteins

The mature protein chain is the amino acid sequence that starts immediately downstream of the predicted signal peptidase cleavage site and extends to the residue immediately upstream of the transcript's stop codon. The relative molecular mass (M_r), isoelectric point (pI) and molar absorption coefficient at 280 nm (ϵ_{280}) of the mature chains were calculated by the PeptideSort program in the GCG software package.

	Protein chain (number of residues)				
Protein	Precursor	Mature	M_r	pI	ϵ_{280}
WAP1	123	99	10874.40	8.06	1760
WAP6	86	61	7073.23	6.92	1490
WAP8	241	203	23469.18	7.86	16120
WAP9	89	66	7953.18	6.88	21150
WAP10	79	59	6854.95	8.23	4080
WAP11	84	59	7166.27	8.42	22430
WAP12	73	53	6221.33	8.23	2800
	89	89	10208.86	8.57	9830
WAP13	93	71	7978.20	7.81	1520
WAP14	231	207	22079.51	7.22	6650

structure. From this position, it stabilizes the protease-binding loop by forming a disulphide bond with the cysteine residue in this loop. Therefore it is likely that WAP6 carries a free cysteine residue in the reactive loop, which might also be more flexible than in the other WAPs.

Expression of the WAP genes

The tissue distribution of transcripts was analysed by RT-PCR on a panel of RNA isolated from 26 different tissues. Transcripts of the APRT gene were amplified simultaneously with the WAP transcripts as a control of assay performance. Different numbers of PCR cycles were tested and the APRT transcript was clearly seen in all samples using 35 cycles, albeit with varying yield. This number of cycles also produced a signal from the novel transcripts and, as can be seen in Figure 5, all the transcripts are highly expressed in the epididymis, but most genes are also transcribed in the testis, with the trachea as the third most common site of transcription. It is also interesting to note that *WAP9*, *WAP10* and *WAP14* appear to be highly transcribed in the brain. In the case of transcripts from *PI3*, *SLPI*, *HE4*, *WAP12* and *WAP14*, the yield was relatively high in all the tissues tested. By decreasing the number of PCR cycles, the signals from several tissues were lost and, following 25 PCR cycles, *WAP12* was only detected in the epididymis and *WAP14* transcripts were almost exclusively seen in the testis. At this number of cycles, the signals of *PI3*, *SLPI* and *HE4* were strongest in the trachea.

DISCUSSION

The clustering of WAP genes at the locus on chromosome 20 suggests that they evolved by repeated duplications. It is tempting to speculate that an early duplication created two genes that subsequently evolved into the centromeric and telomeric clusters of WAP genes. The low degree of sequence conservation indicates that many of these genes were created early on during evolution. Interestingly, probably none of the new WAP genes is the human orthologue of the originally described mouse WAP of lactating mouse [4]. According to human–mouse homology maps (http://www.ncbi.nlm.nih.gov/Homology/), most of human chromosome 20, including the WAP locus described in the present paper, shares conserved synteny with mouse chromosome 2, whereas the

© 2002 Biochemical Society

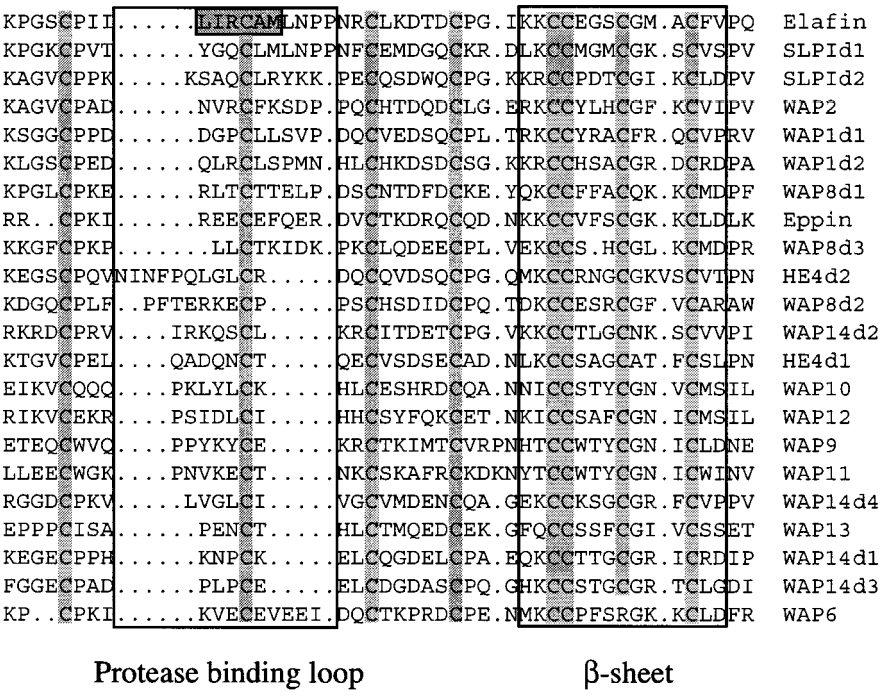

Figure 4 Amino acid sequence alignment

WAP domains expressed by genes from the locus on chromosome 20q12-13.1 were aligned using the computer program PileUp in the GCG package. The alignment was then adjusted manually so that the second cysteine of the WAP motif became residue no. 20. The suffixes d1, d2, d3 and d4 denote the first, second, third and fourth WAP domains respectively, in order from the N-terminus. The conserved cysteines of the WAP motif are shaded. Residues that presumably form the protease-binding loop and the β-sheet at the core of the WAP domain are boxed. The shaded box in the elafin sequence highlights residues P5–P2′ of the primary protease binding loop.

mouse WAP gene is located on chromosome 11 in a region that is equivalent to human chromosome 7p13-12.

Three duplications at the telomeric sub-loci can be traced through the presence of conserved nucleotide sequences. Of these, the similarities between *Eppin* and *WAP6* and between *WAP9* and *WAP11* are of the same magnitude as that previously reported between human semenogelin I and mouse semenoclotin, suggesting that the duplication occurred around the time of the separation of the primate and murine lineages [20]. Likewise, the duplication yielding *WAP10* and *WAP12* probably occurred before the separation of platyrrine and catarrine primates, as the conservation of nucleotides is of the same order as that between semenogelin I genes from man and New World monkeys [21].

Studies of the elafin gene (*PI3*) in the pig has shown that it is duplicated to yield at least three separate genes in this species [22]. There is no indication of such duplications of *PI3* in man, thus the duplications in pig probably occurred after the phylogenetic separation of primates and ungulates.

Previous investigations of the semenogelin genes on chromosome 20 show that they have gone through a rapid and unusual evolution [14,21,23,24]. It appears as if clot proteins of semen have evolved by polypeptide chain extension of more than 10-fold since the phylogenetic separation of primates and rodents. Given the similarity to WAP genes and because many of the WAP genes appear to be phylogenetically old, it is quite possible that the clot proteins evolved from a WAP gene at an early stage of mammalian evolution, perhaps in relation to the development of seminal vesicles.

In contrast with the previously described *WAP2* [17], the novel genes described here do not give rise to transcripts with clearly defined 5′ ends. The reason for this is probably that, unlike *WAP2*, none of them carry a TATA-box in their proximal promoter to direct the site of transcription initiation. In fact, except for *WAP2*, none of the WAP genes at the locus appear to have a classical TATA-box, as *PI3*, *SLPI*, *HE4* and *Eppin* carry what, at best, could be described as TATA-like motifs. Moreover, transcript distribution data for several of the genes indicate that they are ubiquitously transcribed, suggesting that they might be considered as housekeeping genes. This is consistent with the lack of TATA-box, which is a common feature of housekeeping genes.

Although the WAP motif is easy recognized through the cysteine residues that form the conserved disulphides, other residues are poorly conserved in the WAP domains. In part, this probably reflects that most of the genes are very old, and it is also due, in part, to selection for new molecular structures in order to expand the repertoire for protease inhibition. The latter is particularly evident, as shown by the high sequence diversity among the residues that correspond to those of the active loop in elafin. However, the loop also displays size heterogeneity that could indicate that all WAP domains may not manifest protease inhibitory capacity. The proposed protease-binding loop of WAP domains in WAP13 and WAP14 are, for instance, very small in comparison to that of elafin.

The investigation of transcript distribution showed that several WAP genes are transcribed in most of the tissues that were analysed. As the proteins are synthesized as precursor molecules with signal peptide, they are supposedly secreted to the extracellular matrix or to secretory fluids. However, this is probably not true for the deduced product of the WAP12b transcript, which is lacking the normal signal peptide. Thus the product of

© 2002 Biochemical Society

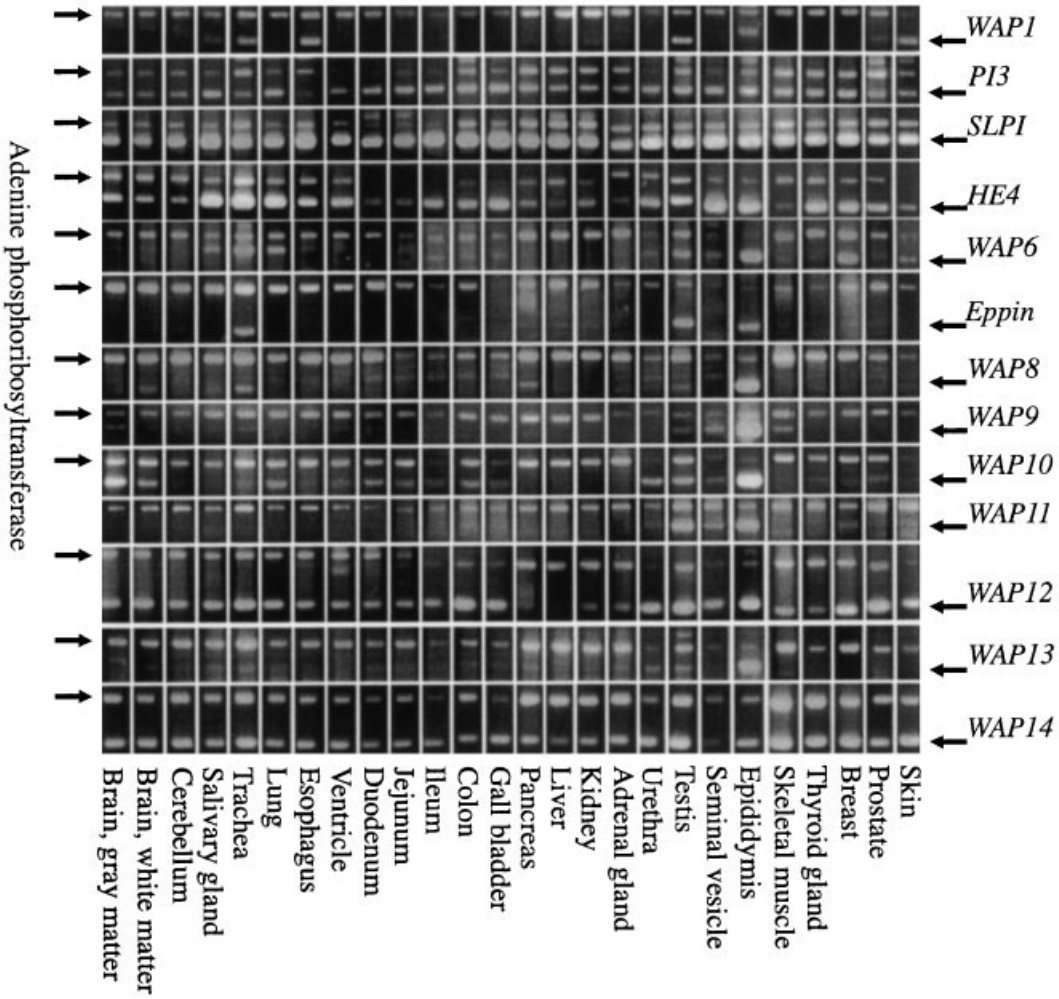

Figure 5 Detection of WAP transcripts by RT-PCR

RNA samples were transcribed into cDNA. Subsequently, transcripts of WAP genes and of the housekeeping gene encoding APRT were simultaneously amplified by 35 cycles of PCR, using specific oligonucleotide primers. Products were analysed by electrophoresis in agarose gels. Mounted photographs are shown of agarose gels stained by ethidium bromide. Tissues used to provide RNA are indicated in the lower part of the Figure.

WAP12b is probably an intracellular protein, unless it is secreted by a mechanism that is independent of signal peptide, as is the case for several transglutaminases [25].

Measurements of SLPI, also known as seminal fluid trypsin–chymotrypsin inhibitor HUSI-1, show high concentrations in both seminal plasma and secretions from bronchi and the parotid gland [26–28]. Many of the recently discovered WAP genes, as reported in the present paper, yield high transcript levels in the epididymis, testis, trachea and salivary gland, suggesting that their gene products might be found at high levels in the same body fluids as SLPI. This also raises the possibility that proteins synthesized from the novel WAP genes could be part of the innate defence against invading micro-organisms, as has been shown for SLPI [29], and that the biological role of the locus is in the regulated inhibition of a wide spectrum of microbial proteolytic enzymes, in order to prevent tissue penetration and infection.

In addition to its activity against micro-organisms, SLPI has also been reported to be an efficient and important inhibitor of endogenous proteases such as elastase and cathepsin G [6]. Endogenous proteases could therefore be a second target of WAPs. It is not yet possible to predict which proteases, other than elastase and cathepsin G, can be regulated by WAPs. However, there is a weak, but intriguing, link between the WAP genes on chromosome 20 and a locus on chromosome 19q13.4, encompassing 15 genes of simple serine proteases that are known as kallikreins, because of their structural similarity to tissue kallikrein [30,31]. Like the WAP genes, the genes of the serine proteases are expressed in a wide variety of tissues, showing distributions that overlap, at least in part, with those of the WAP genes, and there is also a seductive concordance in the number of genes. Among the kallikrein genes are *KLK3* [32,33], the gene that gives rise to prostate-specific antigen (PSA), which is the protein that degrades semenogelin I and II in freshly ejaculated semen [34]. As the semenogelin genes are located within the same locus as the WAP genes from which they probably also evolved, the product of an ancestral gene probably underwent a developmental transformation from inhibitor to substrate. Simultaneously, a glandular kallikrein that was inhibited by the presumed semenogelin ancestor could have been

© 2002 Biochemical Society

duplicated and matured to PSA. Thus, in at least one case, there is a connection between the loci on chromosomes 19 and 20, and forthcoming investigations might also uncover associations between their products.

The excellent technical assistant of Margareta Persson is acknowledged, as is the help of Ingrid Dahlquist and Stefan Strömberg in running the automated DNA sequencer. This investigation was supported by grants from the Swedish Cancer Society (project nos. 4564 and 3555) and the Swedish Research Council (project no. 7903).

REFERENCES

1 Huntington, J. A., Read, R. J. and Carrell, R. W. (2000) Structure of a serpin-protease complex shows inhibition by deformation. Nature (London) **407**, 923–926
2 Ruhlmann, A., Kukla, D., Schwager, P., Bartels, K. and Huber, R. (1973) Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. Crystal structure determination and stereochemistry of the contact region. J. Mol. Biol. **77**, 417–436
3 Bolognesi, M., Gatti, G., Menagatti, E., Guarneri, M., Marquart, M., Papamokos, E. and Huber, R. (1982) Three-dimensional structure of the complex between pancreatic secretory trypsin inhibitor (Kazal type) and trypsinogen at 1.8 Å resolution. Structure solution, crystallographic refinement and preliminary structural interpretation. J. Mol. Biol. **162**, 839–868
4 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
5 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
6 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. U.S.A. **83**, 6692–6696
7 Tsunemi, M., Matsuura, Y., Sakakibara, S. and Katsube, Y. (1996) Crystal structure of an elastase-specific inhibitor elafin complexed with porcine pancreatic elastase determined at 1.9 Å resolution. Biochemistry **35**, 11570–11576
8 Grutter, M. G., Fendrich, G., Huber, R. and Bode, W. (1988) The 2.5 Å X-ray crystal structure of the acid-stable proteinase inhibitor from human mucous secretions analysed in its complex with bovine α-chymotrypsin. EMBO J. **7**, 345–351
9 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
10 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
11 Tomee, J. F., Koeter, G. H., Hiemstra, P. S. and Kauffman, H. F. (1998) Secretory leukoprotease inhibitor: a native antimicrobial protein presenting a new therapeutic option? Thorax **53**, 114–116
12 Legouis, R., Hardelin, J. P., Levilliers, J., Claverie, J. M., Compain, S., Wunderle, V., Millasseau, P., Le Paslier, D., Cohen, D., Caterina, D., Bougueleret, L., Delemarrevandewaal, H., Lutfalla, G., Weissenbach, J. and Petit, C. (1991) The candidate gene for the X-linked Kallmann syndrome encodes a protein related to adhesion molecules. Cell **67**, 423–435
13 Lundwall, A. and Ulvsback, 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
14 Lundwall, A. 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
15 Ulvsback, M., Lazure, C., Lilja, H., Spurr, N. K., Rao, V. V., Loffler, C., Hansmann, I. and Lundwall, A. (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
16 Molhuizen, H. O. F., Zeeuwen, P. L., 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
17 Lundwall, A. A. and Clauss, A. (2002) Identification of a novel protease inhibitor gene that is highly expressed in the prostate. Biochem. Biophys. Res. Commun. **290**, 452–456
18 Richardson, R. T., Sivashanmugam, P., Hall, S. H., Hamil, K. G., Moore, P. A., Ruben, S. M., French, F. S. and O'Rand, M. (2001) Cloning and sequencing of human Eppin: a novel family of protease inhibitors expressed in the epididymis and testis. Gene **270**, 93–102
19 Kirchhoff, C., Habben, I., Ivell, R. and Krull, N. (1991) A major human epididymis-specific cDNA encodes a protein with sequence homology to extracellular proteinase inhibitors. Biol. Reprod. **45**, 350–357
20 Lundwall, A. (1996) The cloning of a rapidly evolving seminal-vesicle-transcribed gene encoding the major clot-forming protein of mouse semen. Eur. J. Biochem. **235**, 424–430
21 Lundwall, A. (1998) The cotton-top tamarin carries an extended semenogelin I gene but no semenogelin II gene. Eur. J. Biochem. **255**, 45–51
22 Tamechika, I., Itakura, M., Saruta, Y., Furukawa, M., Kato, A., Tachibana, S. and Hirose, S. (1996) Accelerated evolution in inhibitor domains of porcine elafin family members. J. Biol. Chem. **271**, 7012–7018
23 Lundwall, A. and Olsson, A. Y. (2001) Semenogelin II gene is replaced by a truncated line 1 repeat in the cotton-top tamarin. Biol. Reprod. **65**, 420–425
24 Ulvsback, M. and Lundwall, A. (1997) Cloning of the semenogelin II gene of the rhesus monkey. Duplications of 360 bp extend the coding region in man, rhesus monkey and baboon. Eur. J. Biochem. **245**, 25–31
25 Aeschlimann, D. and Paulsson, M. (1994) Transglutaminases: protein cross-linking enzymes in tissues and body fluids. Thromb. Haemostasis **71**, 402–415
26 Ohlsson, M., Rosengren, M., Tegner, H. and Ohlsson, K. (1983) Quantification of granulocyte elastase inhibitors in human mixed saliva and in pure parotid secretion. Hoppe-Seyler's Z. Physiol. Chem. **364**, 1323–1328
27 Tegner, H. and Ohlsson, K. (1977) Localization of a low molecular weight protease inhibitor to tracheal and mixillary sinus mucosa. Hoppe-Seyler's Z. Physiol. Chem. **358**, 425–429
28 Fink, E., Jaumann, E., Fritz, H., Ingrisch, H. and Werle, E. (1971) Protease inhibitors in human sperm plasma. Isolation using affinity chromatography and inhibition characteristics. Hoppe-Seyler's Z. Physiol. Chem. **352**, 1591–1594
29 Hiemstra, P. S., Maassen, R. J., Stolk, J., Heinzel-Wieland, R., Steffens, G. J. and Dijkman, J. H. (1996) Antibacterial activity of antileukoprotease. Infect. Immun. **64**, 4520–4524
30 Harvey, T. J., Hooper, J. D., Myers, S. A., Stephenson, S. A., Ashworth, L. K. and Clements, J. A. (2000) Tissue-specific expression patterns and fine mapping of the human kallikrein (KLK) locus on proximal 19q13.4. J. Biol. Chem. **275**, 37397–37406
31 Yousef, G. M., Chang, A., Scorilas, A. and Diamandis, E. P. (2000) Genomic organization of the human kallikrein gene family on chromosome 19q13.3-q13.4. Biochem. Biophys. Res. Commun. **276**, 125–133
32 Riegman, P. H., Vlietstra, R. J., van der Korput, J. A., Romijn, J. C. and Trapman, J. (1989) Characterization of the prostate-specific antigen gene: a novel human kallikrein-like gene. Biochem. Biophys. Res. Commun. **159**, 95–102
33 Lundwall, A. (1989) Characterization of the gene for prostate-specific antigen, a human glandular kallikrein. Biochem. Biophys. Res. Commun. **161**, 1151–1159
34 Lilja, H. (1985) A kallikrein-like serine protease in prostatic fluid cleaves the predominant seminal vesicle protein. J. Clin. Invest. **76**, 1899–1903

Received 5 June 2002/13 August 2002; accepted 2 September 2002
Published as BJ Immediate Publication 2 September 2002, DOI 10.1042/BJ20020869

© 2002 Biochemical Society