Genomic organisation, alternative splicing and primary structure of human matrilin-4

Raimund Wagener*, Birgit Kobbe, Mats Paulsson

Institute for Biochemistry II, Medical Faculty, University of Cologne, Joseph-Stelzmann-Str. 52, D-50931 Cologne, Germany

Received 17 September 1998; received in revised form 5 October 1998

Abstract We have recently cloned a cDNA for mouse matrilin-4. By sequence comparison we identified the 12 kb long human matrilin-4 gene as a part of a high-throughput genomic sequence (HS453C12) in the databases. Additionally we found a human matrilin-4 expressed sequence tag (H54037) in the database that had been mapped to chromosome 20q13.1–2. The gene contains 10 exons and, like the matrilin-1 gene, the human matrilin-4 gene contains an AT-AC intron between the two exons en coding the coiled-coil domain. The cDNA sequence of human matrilin-4 was determined by sequencing of RT-PCR products obtained from mRNA of the human embryonic kidney cell line HEK 293. At the amino acid level it showed an overall sequence identity to the mature mouse matrilin-4 of 91% with a maximum of 97% in the second vWFA-like module. Alternative splicing leads to three different mRNAs. They all encode the putative signal peptide, the two vWFA-like domains and the potential coiled-coil α-helical oligomerisation domain but differ in that either one, two or three EGF-like domains are retained in the mature mRNA. Due to a G to A mutation at the splice donor site of intron C, the third exon encodes an untranslated pseudo-exon specifying the first EGF-like domain when compared to mouse matrilin-4.

Key words: Matrilin; Extracellular matrix; Von Willebrand factor type A; Epidermal growth factor; Coiled-coil α-helix; AT-AC intron

1. Introduction

The matrilin family has to date four members who all share a structure made up of von Willebrand factor A domains, epidermal growth factor-like domains and a coiled coil α-helical module. They accordingly belong to the von Willebrand factor type A-like domain superfamily that consists of proteins with different domain organisation and a variety of functions [1].

Matrilin-1 (previously called cartilage matrix protein or CMP) and matrilin-3 are expressed mainly in cartilage [2–4], while matrilin-2 and -4 occur in a wide variety of extracellular matrices [5,6]. Matrilin-1 forms trimers of identical ellipsoid subunits assembled via their C-terminal extension domains in a coiled-coil α-helix [7] unit and it has also been shown that synthetic peptides corresponding to the C-terminal domain of matrilin-2 assemble into a three-stranded α-helical coiled-coil [8]. Additionally heterotetramers consisting of two subunits of matrilin-1 and two of matrilin-3 have recently been described [9]. Matrilin-1 is tightly associated with cartilage proteoglycans [10,11] as well as being a component of both collagen-dependent [12] and collagen-independent fibrils [13] and on the basis of the related structures other matrilins may play similar roles.

The primary structure of matrilin-1 has been deduced from chicken [14], man [15] and mouse [16] cDNA, of matrilin-2 from mouse and partially from human cDNA [5], of matrilin-3 from mouse [3], chicken [4] and human cDNA [17] and of matrilin-4 from mouse cDNA [6]. In all cases the sequence is highly conserved between species, with the vWFA-like domain being most conserved and the coiled-coil the least conserved domain.

Only the gene structures of matrilin-1 for chicken [18], man [19] and mouse [20] are known and show a similar overall structure. In all three species the codon phases of the introns are conserved and the splice sites show evolutionarily well preserved nucleotide sequences [20]. In the chicken matrilin-1 gene the existence of the minor class of AT-AC introns was shown for the first time [18]. The unusual intron is conserved in the matrilin-1 genes of man and mouse and it is now well accepted that these introns are spliced by the minor subclass of U12-type spliceosomes [21]

Here we report on the organisation of the human matrilin-4 gene and the primary structure of human matrilin-4.

2. Materials and methods

2.1. RNA preparation, RT-PCR and genomic PCR

Total RNA from the lung of a 65-year-old female undergoing lung surgery was prepared by the guanidinium isothiocyanate method. Poly(A^+) RNA was prepared from the total lung RNA and directly from HEK 293 cells and WI-26 human fibroblasts with the QuickPrep Micro mRNA Purification kit (Pharmacia). Human placenta Poly-(A^+) RNA was taken from the Marathon cDNA Amplification kit (Clontech). Reverse transcription using an oligo dT primer was carried out with Superscript II (BRL), the following PCR was carried out with Amplitaq (Perkin Elmer) on two overlapping fragments using primers h41 or h420 and h421 for the N-terminal and h414 and h423 for the C-terminal fragments. PCR conditions after an initial incubation at 94°C for 3 min were 40 cycles 1 min 94°C, 1 min T_m−5°C (4+2 rule), 1 min 72°C with the standard protocol and additional 10% DMSO.

Genomic PCR was performed with 100 ng human or mouse genomic DNA per reaction using the Expand High Fidelity PCR system (Boehringer Mannheim) following the manufacturer's recommendations.

*Corresponding author. Fax: (49) (221) 4786977.
E-mail: raimund.wagener@uni-koeln.de

Abbreviations: vWFA, von Willebrand factor type A; EGF, epidermal growth factor; EST, expressed sequence tag; RT-PCR, reverse transcription polymerase chain reaction; HTGS, high-throughput genomic sequence

The nucleotide sequences reported in this paper have been submitted to the EMBL nucleotide sequence database with accession numbers AJ010984 and AJ007581.

PII: S0014-5793(98)01293-9

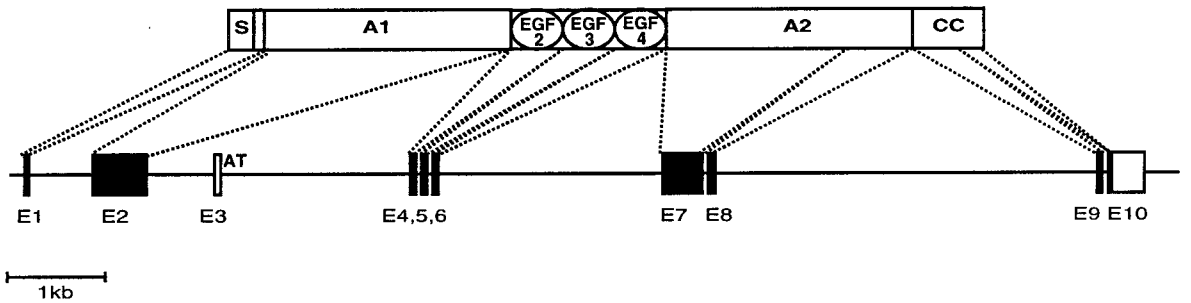

Fig. 1. Organisation of the human matrilin-4 gene. Exon-intron organisation of the human gene for matrilin-4 and the assignment of exons to the structural domains of human matrilin-4. S, signal peptide; A1 and A2, vWFA-like domains; CC, coiled-coil oligomerisation domain. The pseudo-EGF1 exon 3 and the 3′ UTR in exon 10 are represented as open boxes.

Oligonucleotides used for human matrilin-4 were: h425 5′-GTTGCCAGCACCAATGTGTCAATG-3′ (nt 76–99); h426 5′-CAAACAGCTCTTGTTATCTGCTGC-3′ (nt 144–167). The numbering of the nucleotides is according to Fig. 2. h41 5′-ACACCTCCATTCTCTGTCC-3′ (nt 1–19); h420 5′-ATGAGAGGCCTTCTTTGCTGGCCCG-3′ (nt 29–53); h414 5′-GCTTTGTACTCCAGCAGGAC-3′ (nt 759–778); h421 5′-CCTCGCTCACACACTGGAACTC-3′ (nt 974–953); h423 5′-ATCCAACAAAGAACTGAGCGCAG-3′ (nt 2048–2026). The numbering of the nucleotides is according to Fig. 3. Oligonucleotides used for mouse matrilin-4 were: m42 5′-AGTCCTTCGATCTCATCC-3′ (nt 685–702); m414 5′-CATTCTACTGCGCCTGCAACTC-3′ (nt 802–823); m416 5′-GCCAAACAATTCTTGTTATCTG-3′ (nt 862–841); m415 5′-CGACAGAAATACGAGTCCACGGAG-3′ (nt 937–914). The numbering of the nucleotides is according to [6].

2.2. DNA sequencing

DNA sequencing was performed on PCR products using the ABI Prism Big Dye Terminator Cycle Sequencing Ready Reaction kit and the products were resolved on an ABI Prism 377 automated sequencer (Perkin-Elmer/Applied Biosystems). PCR products were sequenced directly at least twice in both directions to exclude PCR artifacts. Nucleotide sequence analysis was performed with the programs of the GCG package [22].

3. Results

3.1. Human matrilin-4 gene

By screening the databases with the recently determined mouse matrilin-4 cDNA sequence as a query we identified the entire human matrilin-4 gene on a contig of a phase I HTGS [23] (high-throughput genomic sequence, clone HS453C12, accession number AL021578). The contig is located on chromosome 20 and the gene spans about 12 kb consisting of at least 10 exons (Fig. 1, Table 1). The putative existence of another intron in the 5′ untranslated region has to be resolved when the transcription start is determined. Supposing that no more introns exist, the putative promoter contains by computer analysis [24] a TATA motif-like sequence, TGTAAAT, 75 bases upstream of the translation start, whereas no characteristic CCAAT box is found. Exons were identified by flanking consensus splice signals and by comparison with the cDNA sequence. The first exon (E1) encodes the signal peptide and a short sequence of 8 amino acid residues N-terminal of the first vWFA domain. The first vWFA domain (E2) and the four EGF-like domains (E3, E4, E5 and E6) are each encoded by single exons, whereas the second vWFA domain (E7 and E8) and the coiled-coil domain (E9 and E10) are represented by two exons each. With the exception of the single coding exon for the vWFA1 domain the exon-intron organisation of the human matrilin-4 gene shows some similarities to the modular organisation of the genes for matrilin-1 of man [19], mouse [20] and chicken [18]. The most interesting common feature is the unusual intron (I) that divides the coiled-coil domain and belongs in the case of matrilin-1 to a subgroup of introns that are spliced by the U12-type spliceosome [21]. The 5′ consensus splice site (ATATCCTTT) for U12-type introns is completely conserved in the human matrilin-4 gene. The intron also contains a putative branch site (CTCCTTAACCGC) 10 nucleotides upstream of the 3′ splice site that is highly homologous to the consensus branch site (TTCCTTRACYCY) of U12-type AT-AC introns [21].

Table 1
Splice junction sites of the human matrilin-4 gene derived from HTGS HS453C12

Exon	Size (bp)	Splice donor	Intron	Size (bp)	Splice acceptor	Codon phase	Amino acid interrupted
1	73[a]	GACAGgtgaac	A	711	ctccccgcagGTC	I	G (25)
2	570	GTGTGgtgagt	B/C	2784	tcttcctcagCCA	I	A (215)
4	123	CAGGGgtgagc	D	96	ctcctgccagCCA	I	A (256)
5	123	TCAGGgtgagt	E	92	ctcctttcagTCC	I	V (297)
6	123	CAACCgtgagt	F	2303	ggctgtgcagGGT	I	R (338)
7	414	GGAAGgtgggc	G	99	ggccttgcagGCA	I	G (476)
8	153	TCCAGgtgagc	H	3887	cggcttgcagAGG	I	E (527)
9	108	GAACC**at**atcc	I	82	aaccgctc**ac**TGG	I	L (563)
10	56[a]						
Consensus for U2-type GT-AG introns [21]		KAGgtragt			yyyyyyncagG		

Exon sequences are in capital letters, intron sequences are in lower case letters. The AT-AC splice site is in bold.
[a]The length of the 5′ and 3′ UTRs in exon 1 and exon 10 is not included.

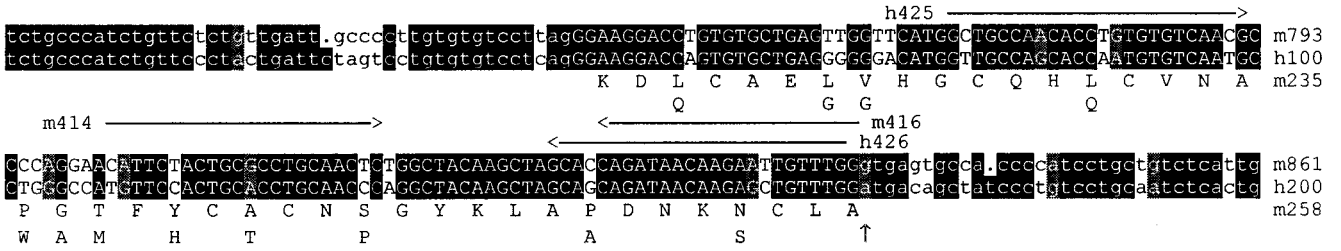

Fig. 2. Comparison of exon 3 of human and mouse matrilin-4 and flanking intron sequences. Intron sequences are given in lower case, exon sequences in upper case letters. Mouse sequences are given in the upper line, human sequences in the lower line. For human only the amino acids deviations are given. The length and the orientation of synthetic oligonucleotides used to prime the PCR are indicated by horizontal arrows. The position of the corresponding mouse primers m42 (nt 685–702) for the m416/m42 PCR and m415 (nt 937–914) for m414/m415 PCR are not given and can be derived from [6]. The numbering of the mouse exon sequence is according to [6]. The human splice acceptor site mutation is indicated by a vertical arrow.

Interestingly, we found in the human matrilin-4 gene an AT-AG intron (C) between exon 3 and 4 that has not been reported to be excised regularly [21]. In addition a GT-AT transition of 5′ splice donor sites is described in a number of genetic disorders caused by skipping of the preceding exon or the use of alternative splice sites leading to frameshift mutations or loss of essential functional domains. To exclude sequencing errors in the HTGS phase I sequence we amplified the exon-intron border by PCR and confirmed the sequence by direct sequencing of the PCR product (data not shown). Additionally, using primer pairs m42/m416 and m414/415 we determined the sequence of the corresponding part of the mouse matrilin-4 gene from a genomic PCR product and found the unmutated canonical GT splice donor site of U2-type GT-AG introns (Fig. 2). In the human matrilin-4 gene we identified a putative branch site sequence (CTGAC) 18 nucleotides upstream of the 3′ splice site that fits perfectly to the consensus sequence for the putative branch site of U2-type GT-AG introns (CTRAY) [21]. The 5′ splice site (ATGAGA) of intron C is homologous to the 5′ splice site of U2-type GT-AG introns (GTRAGT) and has no homology with the 5′ splice sites of U12-type AT-AC or U12-type GT-AG introns [21], clearly indicating that the intron is a mutated U2-type GT-AG intron.

3.2. Matrilin-4 cDNA

By RT-PCR we could detect expression of matrilin-4 in the human embryonic kidney cell line HEK 293, adult lung, placenta and WI-26 human fibroblasts (Fig. 4), whereas we were not able to detect expression on poly(A^+) Northern blots due to a low expression level in the examined cells or tissues (not shown). To date three human matrilin-4 EST sequences, AA992948, AI038302 and H54037, have been published. They are from total human foetuses, 8–9 weeks and 6 weeks old, respectively. EST H54037 is characterised as a rarely expressed cDNA and is mapped to q13.1–2 of chromosome 20 with in situ hybridisation using a radiolabelled probe [25].

To determine the complete coding region of the human matrilin-4 cDNA we performed RT-PCR with primers designed from the putative exons on the genomic sequence. RT-PCR products derived from the HEK 293 cell mRNA were sequenced directly (Fig. 3). Only two differences with the exon sequences of HS453C12 were detected. In exon 2 a silent substitution (A-G, nt 376, Fig. 3) and a single base transversion (C-A, nt 518, Fig. 3) resulting in an arginine-to-serine substitution are present only in HEK 293 cells and placenta, whereas sequencing of the corresponding human lung RT-PCR products confirmed the HTGS. The RT-PCR resulted in three differentially spliced human matrilin-4 cDNAs that were all different in the modular structure with respect to the two differentially spliced mouse transcripts (Fig. 5). All cDNAs contained the same sequences for the signal peptide, the vWFA-like modules and the coiled-coil module, but differed in the number of EGF-like modules. Full-length human matrilin-4 could be detected neither in HEK 293 cells nor in lung, placenta and WI-26 human fibroblasts where all the other matrilin-4 RT-PCR products could be detected. Mainly RT-PCR products containing EGF-like modules 2, 3 and 4 or products containing EGF-like modules 3 and 4 and to a much lesser content containing only EGF-like module 4 were detected (Fig. 4A).

In the second splice variant the codons $G^{671}C^{672}C^{673}$ of aa 215 (Ala) and $G^{794}C^{795}C^{796}$ of aa 256 (Ala) are fused to a resulting new GCC codon (Ala) leading to the transition nucleotide sequence $G^{671}C^{795}C^{796}$ and the corresponding amino acid sequence $C^{214}A\ I^{257}$. In the shortest splice variant the codons $G^{671}C^{672}C^{673}$ of aa 215 (Ala) and $G^{916}T^{917}C^{918}$ of aa 297 (Val) are fused to a resulting new GTC codon (Val) leading to the transition nucleotide sequence $G^{671}C^{917}C^{918}$ and the corresponding amino acid sequence $C^{214}V\ R^{298}$. Alternatively spliced transcripts lacking the domain corresponding to vWFA1 of mouse matrilin-4 [6] were not detected in human matrilin-4. To unequivocally show the deficiency of the first EGF-like module we performed genomic and RT-PCR on HEK 293 cell mRNA with EGF1 module-specific primers and primers from the neighbouring vWFA1 and EGF4 modules in both directions. Only genomic PCR products but no RT-PCR products could be detected (Fig. 4A,B), clearly indicating that exon 3 is a pseudo-exon. While it encodes the first EGF-like domain, it is not retained in the mature mRNA.

The overall identity to the mature mouse matrilin-4 is 91% and the most conserved parts are the vWFA modules with an identity of 92% for vWFA1 and 97% for vWFA2. The EGF-like modules have an identity of 71% for the pseudo-EGF1 and 80%, 85% and 90% for EGF2, -3 and -4 respectively. The identity of the coiled-coil domain is 87% but the heptad positions are perfectly conserved. The position and the spacing between the cysteine residues and the MIDAS motif [26] are identical in the mature proteins. The signal peptide of the human matrilin-4 lacks two amino acids but the predicted

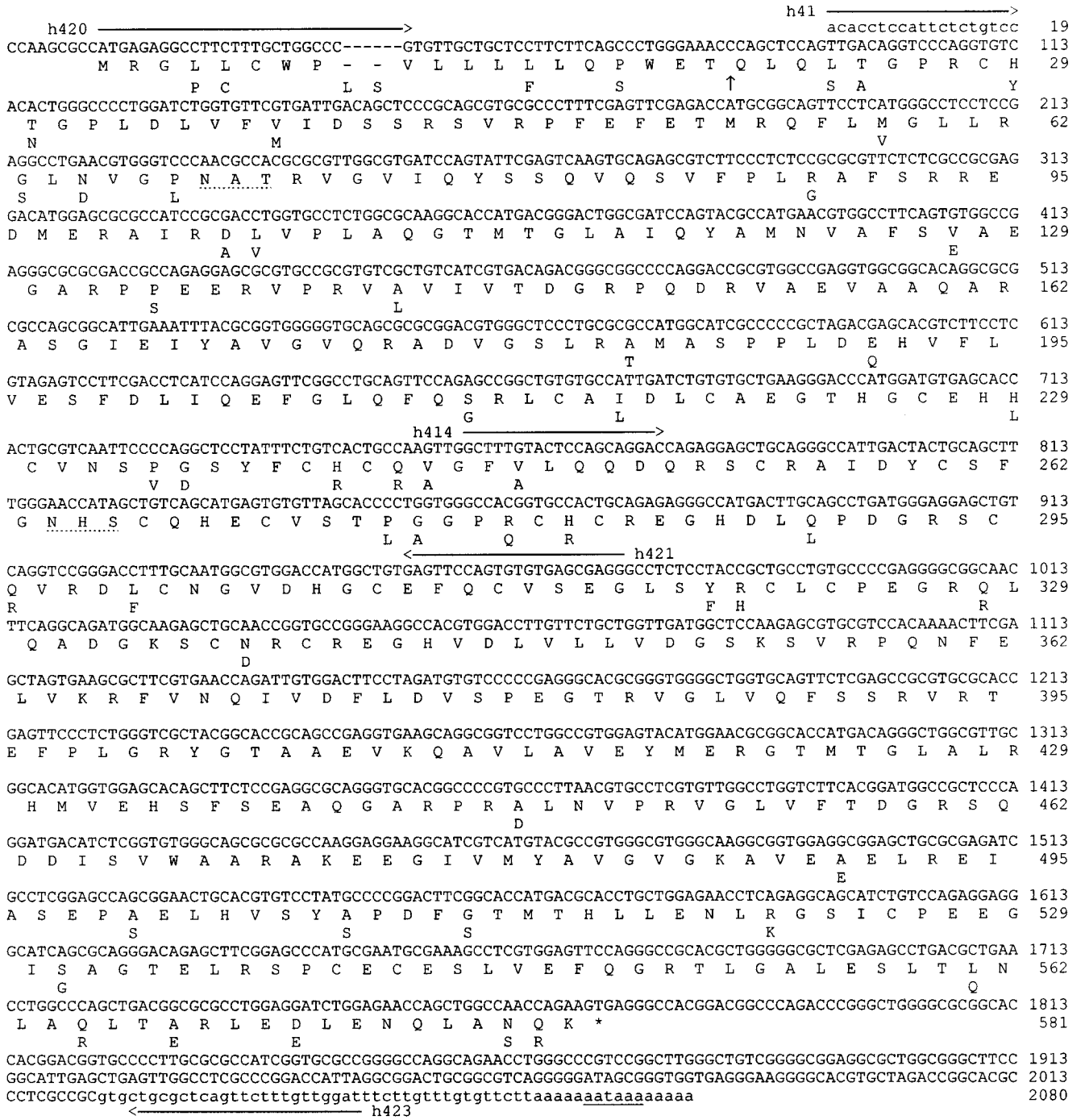

Fig. 3. cDNA and deduced amino acid sequence of human matrilin-4. The deviations in the mouse matrilin-4 amino acid sequence are given below. nt 1–19 and 2023–2080 are from HS453C12 and are given in lower case letters. The length and the orientation of synthetic oligonucleotides used to prime the PCR are indicated by horizontal arrows. The potential signal peptide cleavage site is indicated by a vertical arrow. The potential *N*-glycosylation sites are underlined with a dotted line, the polyadenylation signal with a solid line.

signal peptide cleavage site [27] is identical in the human and mouse matrilin-4. The three different splice variants yield mature protein products with predicted M_rs of 61 774, 57 255 and 52 784 respectively. Both *N*-glycosylation sites are conserved in the human sequence (Asn^{69}-Ala-Thr and Asn^{264}-His-Ser). A polyadenylation signal 296 bp downstream of the stop codon could be identified on the genomic sequence (Fig. 3).

4. Discussion

We have identified the human matrilin-4 gene on an HTGS in the databases and with the sequence information we were able to determine the primary structure by sequencing of RT-PCR products. The overall high identity clearly demonstrated that it is the human homologue of mouse matrilin-4. The temporal and spatial expression pattern is also similar. Never-

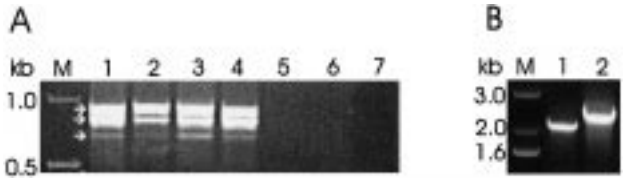

Fig. 4. Distribution of human matrilin-4 mRNA and evidence that exon 3 is a pseudo-exon. A: Comparative RT-PCR analysis. Lanes 1–4: placenta, lung, WI-26 human fibroblasts and HEK 293 cells, respectively. Primers for lanes 1–4 were h420/h421. Lanes 5+6: HEK 293 cells, primers h420/h426 and h421/h425, respectively (see Section 2). Lane 7: negative control with primers h420/h421. Arrows indicate the three differentially spliced RT-PCR products. B: Genomic PCR analysis of exon 3. Lane 1: primers h420/h426. Lane 2: primers h421/h425.

theless the human matrilin-4 has striking differences in modular composition of the different alternatively spliced gene products (Fig. 5). The first EGF-like domain is not expressed due to a GT-AT mutation in the 5′ splice donor site of intron C, although exon 3 encoding the first EGF is still present as a pseudo-exon in the human matrilin-4 gene (Fig. 2). Pseudo-exons are not unusual and can be found for example in human αA-crystallin [28] or rat aggrecan, where a heavily mutated non-expressed EGF-like domain is present [29]. It is interesting to see how substitution of one single nucleotide in evolution can result in the loss of a protein module. The homology to the corresponding expressed mouse matrilin-4 sequence is still 70% at the amino acid level, but clearly lower than the identities of the expressed EGF-like domains, indicating that the substitution occurred very recently in evolution.

A splice product lacking the first vWFA-like domain as described in mouse matrilin-4 [6] has so far not been detected in humans, but our access to human material from different tissues and developmental stages has been limited. On the other hand, we have not detected mouse matrilin-4 mRNA lacking any of the EGF-like domains in a broad collection of tissues at different developmental stages. It may be that this difference in alternative splice pattern is the effect of the evolutionary fine tuning of the protein function in two different mammalian species. As the EGF modules are believed to have a spacing function alternative splicing will provide different

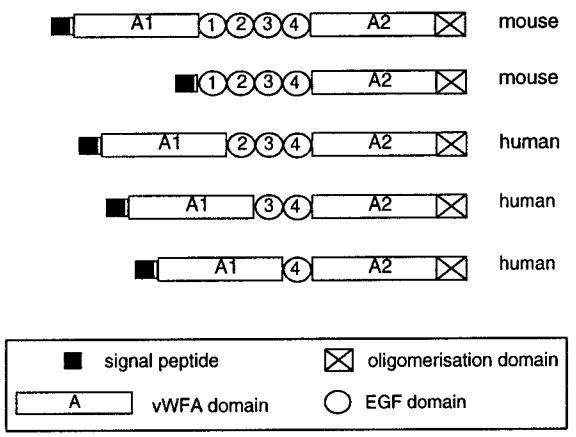

Fig. 5. Splice variants of matrilin-4. The domain structures of the splice variants detected in mouse or man are depicted.

distances between the vWFA-like domains, which are likely to carry important ligand binding sites.

The matrilin-4 gene, like the matrilin-1 gene [18–20], contains a U12-type AT-AC intron that separates the two exons encoding the coiled-coil domain. This is further strong evidence for a common origin of the members of the matrilin family in addition to the high structural homology of the proteins. It is speculated that the removal of AT-AC introns is rate-limiting for the maturation of their host mRNAs and thereby contributes importantly to gene regulation [30]. It is not unlikely that also the matrilin-2 and -3 genes contain a U12-type AT-AC intron and the investigation of the matrilins can lead to further insights into the function of this novel type of spliceosome.

Acknowledgements: This work was supported by Köln Fortune program/Faculty of Medicine, University of Cologne. We thank Dr. Krüger from the Department of Surgery, University Clinic, Cologne for the lung specimen and Ferenc Deak, Tibor Rauch, Andreas Klatt and Patric Nitsche for valuable discussions.

References

[1] Colombatti, A. and Bonaldo, P. (1991) Blood 77, 2305–2315.
[2] Paulsson, M. and Heinegård, D. (1982) Biochem. J. 207, 207–213.
[3] Wagener, R., Kobbe, B. and Paulsson, M. (1997) FEBS Lett. 413, 129–134.
[4] Belluoccio, D. and Trueb, B. (1997) FEBS Lett. 415, 212–216.
[5] Deák, F., Piecha, D., Bachrati, C., Paulsson, M. and Kiss, I. (1997) J. Biol. Chem. 272, 9268–9274.
[6] Wagener, R., Kobbe, B. and Paulsson, M. (1998) FEBS Lett. (in press).
[7] Hauser, N. and Paulsson, M. (1994) J. Biol. Chem. 269, 25747–25753.
[8] Pan, O.H. and Beck, K. (1998) J. Biol. Chem. 273, 14205–14209.
[9] Wu, J.J. and Eyre, D.R. (1998) J. Biol. Chem. 273, 17433–17438.
[10] Paulsson, M. and Heinegård, D. (1979) Biochem. J. 183, 539–545.
[11] Hauser, N., Paulsson, M., Heinegård, D. and Mörgelin, M. (1996) J. Biol. Chem. 271, 32247–32252.
[12] Winterbottom, N., Tondravi, M.M., Harrington, T.L., Klier, F.G., Vertel, B.M. and Goetinck, P.F. (1992) Dev. Dyn. 193, 266–276.
[13] Chen, Q., Johnson, D.M., Haudenschild, D.R., Tondravi, M.M. and Goetinck, P.F. (1995) Mol. Biol. Cell 6, 1743–1753.
[14] Argraves, W.S., Deak, F., Sparks, K.J., Kiss, I. and Goetinck, P.F. (1987) Proc. Natl. Acad. Sci. USA 84, 464–468.
[15] Jenkins, R.N., Osborne-Lawrence, S.I., Sinclair, A.K., Eddy, R.L., Byers, M.G., Shows, T.B. and Duby, A.D. (1990) J. Biol. Chem. 265, 19624–19631.
[16] Aszódi, A., Hauser, N., Studer, D., Paulsson, M., Hiripi, L. and Bösze, Z. (1996) Eur. J. Biochem. 236, 970–977.
[17] Belluoccio, D., Schenke, T., Balci, A. and Trueb, B. (1998) Genomics (in press).
[18] Kiss, I., Deák, F., Holloway, R.G., Delius, H., Mebust, K.A., Frimberger, E., Argraves, W.S., Tsonis, P.A., Winterbottom, N. and Goetinck, P.F. (1989) J. Biol. Chem. 264, 8126–8134.
[19] Jenkins, R.N., Osborne-Lawrence, S.L., Sinclair, A.K., Eddy Jr., R.L., Byers, M.G., Shows, T.B. and Duby, A.D. (1990) J. Biol. Chem. 265, 19624–19631.
[20] Aszódi, A., Beier, D.R., Hiripi, L., Bösze, Z. and Fässler, R. (1998) Matrix Biol. 9, 563–573.
[21] Sharp, P.A. and Burge, C.B. (1997) Cell 91, 875–879.
[22] Genetics Computer Group (1997) Program Manual for the Wisconsin Package, Version 9.1, Genetics Computer Group, Madison, WI.
[23] Ouellette, B.F. and Boguski, M.S. (1997) Genome Res. 7, 952–955.

[24] Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) Nucleic Acids Res. 23, 4878–4884.
[25] Jay, P., Diriong, S., Taviaux, S., Roeckel, N., Mattei, M.G., Audit, M., Berge-Lefranc, J.L., Fontes, M. and Berta, P. (1997) Genomics 39, 104–108.
[26] Lee, J.-O., Rieu, P., Arnaout, M. and Liddington, R. (1995) Cell 80, 631–638.
[27] Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Protein Eng. 10, 1–6.
[28] Jaworski, C.J. and Piatigorsky, J. (1989) Nature 337, 752–754.
[29] Doege, K.J., Garrison, K., Coulter, S.N. and Yamada, Y.J. (1994) J. Biol. Chem. 269, 29232–29240.
[30] Tarn, W.Y. and Steitz, J.A. (1997) Trends Biochem. Sci. 22, 132–137.