Short Note

MATRIX BIOLOGY

Characterization of the Human Extracellular Matrix Protein 1 Gene on Chromosome 1q21

MAUREEN R. JOHNSON*, DOUGLAS J. WILKIN*, HANS L. VOS†, ROSA ISELA ORTIZ DE LUNA‡, ANINDYA M. DEHEJIA§, MIHAEL H. POLYMEROPOULOS§ and CLAIR A. FRANCOMANO*

* Medical Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA,
† Netherlands Cancer Institute, Amsterdam, The Netherlands,
‡ Hospital Infantil de Mexico "Federico Gomez", Mexico City, Mexico and
§ Laboratory of Genetic Disease Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA.

Abstract

Ecm1, the mouse gene encoding extracellular matrix protein 1, is highly expressed in bone and cartilage as well as in osteogenic, preosteoblastic and chondroblastic cell lines. Ecml was recently localized to a chromosomal region in mouse syntenic to human chromosome 1q21, establishing this gene as a prime candidate gene for pycnodysostosis, a rare, autosomal recessive sclerosing skeletal dysplasia. Shortly thereafter, it was determined that cathepsin K is the pycnodysostosis gene. We now report the radiation hybrid mapping of human ECM1 to 1q21, and the gene structure and coding sequence of human ECM1.

Key words: bone, cartilage, chromosome 1q21, extracellular matrix protein 1.

Introduction

The murine gene extracellular matrix protein 1 (ECM-1) encodes a novel secretory protein with structural similarity to serum albumin family proteins and Endo 16, a calcium binding protein from sea urchin (Bhalerao et al., 1995). The protein was originally found to be secreted from the osteogenic cell line MN7, established from bone marrow stroma of the adult mouse (Bhalerao et al., 1995). Strong expression of the Ecml gene was seen in skin and cartilage containing tissues, as well as in preosteoblastic and chondroblastic cell lines. Ecm-1 maps to a region in the mouse that is syntenic to human chromosome 1q21. This chromosomal localization in mouse led us to consider ECM-1 briefly as a candidate gene for the human autosomal recessive sclerosing skeletal dysplasia, pyknodsystosis. This line of inquiry was abandoned when the gene encoding cathepsin K was found to be the causative gene for pyknodsystosis (Gelb et al., 1996; Johnson et al., 1996). We now report the human chromosomal localization of ECM1[1] and characterize its genomic structure and coding sequence.

[1] HGMW-approved symbol for the human homolog of the mouse Ecml gene is ECM1.

[2] D. Slonim, L. Stein, L. Kruglyak and E. Lander, unpublished software.

Matrix Biology Vol. 16/1997, pp. 289–292

Methods and Results

Radiation hybrid mapping of human ECM-1

Mapping of ECM1 was performed using the Genebridge 4 radiation hybrid panel (Walter et al., 1994). Oligonucleotide primers used for PCR were 5′-GACGTAACATCTGGCGAGAC-3′ (forward) and 5′-CATCTGACCCTCTAGGGCTCTG-3′ (reverse), which amplified a 226 bp fragment from the 3′ end of the ECM1 gene. Statistical analysis of the data was performed using RHMAPPER software[2]. The data vector for ECM-1 was 020121200122101001100000120110011100001010121010010101001200000000001201011000120102000000211, localizing it between markers WI-497 and D1S305 on chromosome 1, 15.6 cR from WI-497 and 11.4 cR from D1S305 (lod > 3.0). This localization is consistent with the genetic mapping on 1q21.

Human ECM1 gene structure and sequence analysis

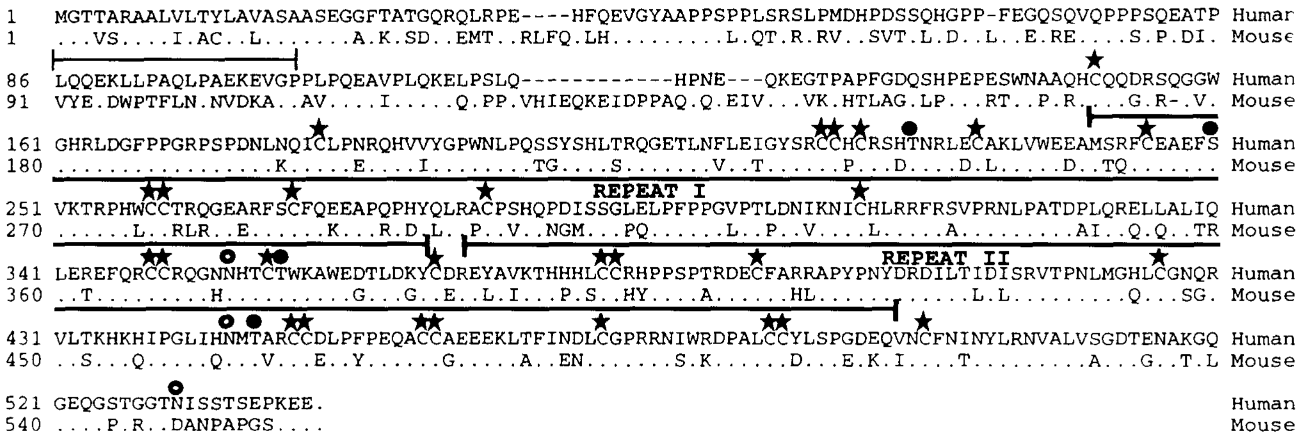
The mouse Ecm1 cDNA sequence (Bhalerao et al., 1995) was used to perform a homology search to identify human expressed sequence tagged sites (ESTs) that correspond to the human homolog of the mouse Ecm1 cDNA. A contig of these sequences (GenBank accession numbers H66471, N99800, R83413, T84826, N79484, N71317, N71368, R83319, H90991, T71394, R62857, H68980, and H66728) was constructed and aligned to the mouse sequence using the MegAlign sequence analysis software program (DNASTAR, Inc.). Two large gaps in the human sequence were identified corresponding to sequence missing in the EST database, one gap in the 5′ end of the gene encompassing 180 bp, and a second gap spanning approximately 90 bp. Based on the human EST sequence, primers were designed to amplify and sequence the entire ECM1 coding sequence and intron/exon boundaries of genomic DNA. Oligonucleotide primer sequences and PCR conditions are available upon request. Sequences for intron/exon boundaries were established where the genomic sequence diverged from the mouse/human cDNA alignment (Shapiro and Senapathy, 1987). Sequencing the first gap revealed intron sequences at both the 5′ and the 3′ end of the gap, resulting in the necessity to sequence this entire genomic region. In addition, the I.M.A.G.E. consortium clone ID #229703 (Lennon et al., 1996; clone obtained from Research Genetics), which, based on insert size, contains full length human ECM1 cDNA, was sequenced to further define intron/exon boundaries. Sequencing the second gap region was straightforward, since this region did not contain an intron.

The human ECM1 coding sequence encompassed 1.6 kb and showed 79% identity when compared to the mouse coding sequence. Sequence data have been deposited in the GenBank database and may be retrieved using the accession numbers U65932 through U65938. Since the PCR products included intron sequence, size

Table I. Nucleotide sequence of the intron/exon boundaries of human *ECM1*. The first exon is denoted as the one with the initiating methionine residue. Invariant donor and acceptor splice site nucleotides are underlined. The entire sequence of intron 2 and intron 4 are shown. The size of the exons and intervening sequence (IVS) are denoted at the right, where * indicates approximated size.

Intron/Exon Borders	Exon	Size (bp)	IVS	Size (bp)
aggacccacctctgagtgtccagtggtcagttgccccagg/ATG Exon1/ gtgagttgggggatcagcacttaggagggggtctgggctt ...	1	70	1	1400*
...gccctcccagtggcccctgacttgcccttcttccctccag/Exon2/ gtaagagtttgggggagcagcatgggattgggactccagg	2	51	2	79
aggcactgtgggctctgatgtctcccctcttgcttctag/Exon3/ gtaaggtcaccatcccatgccctctcagtgaccctccag ...	3	102	3	159
...ggagaaagggtgggctgctcacacattccccttctatag/Exon4/ gtgagcgcttgcccaccctccctcacctctatcccactatggatcctgttgacaccagg	4	81	4	99
ctgatccctgctccttggtgccctcacccctatcttgcag/Exon5/ gtaagcagctccctctcttctttacccacctttacctcat ...	5	81	5	693
...cttcacatgtccccgcttcccactgttttccccattccag/Exon6/ gtaaggttgggttcttgatgccggggggtgtcctttaacc ...	6	323	6	400*
...ggaatgtggaaagtgggctgatcctcccctcttgctctag/Exon7/ gtaagtgggcgtcccagcctccctgagagcctgtttgcct ...	7	375	7	600*
...caagtgtccagcttctgacttccctctctctggtccacag/Exon8/ gtaagttgcctaatccttccccactetettcctttcccga ...	8	221	8	153
...cccaaagaccctaacccctgccccttcacaccaacatag/Exon9/ gtgagtgtgtggagtctagtctccagaggaatgcaggga ...	9	88	9	450*
...tccccaccccatcatctgtttacttttctcattcatcag/Exon10 TGA	10	231	–	–

A



```
1   MGTTARAALVLTYLAVASAASEGGFTATGQRQLRPE----HFQEVGYAAPPSPPLSRSLPMDHPDSSQHGPP-FEGQSQVQPPPSQEATP Humar
1   ...VS....I.AC..L.......A.K.SD..EMT..RLFQ.LH.........L.QT.R.RV..SVT.L.D..L..E.RE....S.P.DI. Mouse

86  LQQEKLLPAQLPAEKEVGPPLPQEAVPLQKELPSLQ------------HPNE---QKEGTPAPFGDQSHPEPESWNAAQHCQQDRSQGGW Human
91  VYE.DWPTFLN.NVDKA..AV....I.....Q.PP.VHIEQKEIDPPAQ.Q.EIV...VK.HTLAG.LP...RT..P.R....G.R-.V. Mouse

161 GHRLDGFPPGRPSPDNLNQICLPNRQHVVYGPWNLPQSSYSHLTRQGETLNFLEIGYSRCCHCRSHTNRLECAKLVWEEAMSRFCEAEFS Human
180 ................K.....E....I........TG....S.......V..T......P...D....D.L.....D..TQ....... Mouse

                                                             REPEAT I
251 VKTRPHWCCTRQGEARFSCFQEEAPQPHYQLRACPSHQPDISSGLELPFPPGVPTLDNIKNICHLRRFRSVPRNLPATDPLQRELLALIQ Human
270 ......L..RLR..E......K...R.D.L..P..V..NGM...PQ......L..P..V....L.....A.........AI..Q.Q..TR Mouse

                                                                           REPEAT II
341 LEREFQRCCRQGNNHTCTWKAWEDTLDKYCDREYAVKTHHHLCCRHPPSPTRDECFARRAPYPNYDRDILTIDISRVTPNLMGHLCGNQR Human
360 ..T.........H..........G...G..E..L.I...P.S..HY....A......HL............L.L.........Q...SG. Mouse

431 VLTKHKHIPGLIHNMTARCCDLPFPEQACCAEEEKLTFINDLCGPRRNIWRDPALCCYLSPGDEQVNCFNINYLRNVALVSGDTENAKGQ Human
450 ..S...Q.....Q...V...E..Y......G.....A..EN.......S.K......D...E.K.I....T.........A...G..T.L Mouse

521 GEQGSTGGTNISSTSEPKEE.                                                                      Human
540 ....P.R..DANPAPGS....                                                                      Mouse
```

B

sequences	amino acid numbers		
	human	mouse	identity (%)
signal peptide sequence	1–19	1–19	68
cysteine-free region	20–150	20–170	45
repeat I	151–279	170–298	79
spacer	280–300	299–319	62
repeat II	301–405	320–424	76
remainder	406–541	425–560	73
overall	1–541	1–560	69

Fig. 1. (A) Deduced amino acid sequence alignment of the human and mouse ECM1 proteins. Alignment was performed using the MegAlign sequence analysis software program (DNASTAR, Inc.). Numbering begins with the initiating methionine codon. Dots denote amino acid homology, and dashes represent deleted residues. Conserved cysteine residues are denoted by stars. Open circles represent the three potential N-glycosylation sites, and closed circles represent potential protein kinase C phosphorylation sites. The putative signal peptide is denoted by a thin underline. Repeat regions I and II are underlined in bold. (B) Conservation between human and mouse ECM1. The putative signal peptide sequences (residues 1–19) are 68% conserved. This region is followed by a cysteine-free region spanning residues 20–150 in humans and residues 20–170 in mouse, and is not highly conserved between species. Twenty amino acids are deleted in the human sequence when compared to mouse sequence in this region, resulting in only 45% identity between species at the protein level. Following this sequence, the remainder of the protein can be subdivided into a region containing cysteine residues and also a repeated sequence. Repeat region I (residues 151–279 in humans and 170–298 in mouse) is 79% identical between species, and repeat II (residues 301–405 in humans and 320–424 in mouse) is 76% identical. The spacer region between Repeat I and Repeat II, from residues 280–300 in humans and residues 299–319 in mouse, is 62% identical, with the remainder of the protein 73% identical. The overall conservation between the human and mouse ECM1 proteins is 69%.

determination of these products demonstrated that the total genomic sequence is approximately 5.6 kb in length, excluding the 5′ and 3′ untranslated regions (data not shown). Genomic sequence divergence from the cDNA sequence and the identification of sequences for canonical donor and acceptor splice sites (Shapiro and Senapathy, 1987) determined that the gene consists of ten exons and nine introns (Table I). Deduced amino acid sequence showed 69% identity to the mouse deduced amino acid sequence (Fig. 1).

Discussion

The recent report of the characterization of the mouse Ecm1 cDNA (Bhalerao et al., 1995) and its localization to a region of the mouse chromosome that is syntenic to human chromosome 1q21 suggested that ECM1 may be the gene disrupted in pycnodysostosis. Mouse Ecm1 is a secreted protein that is highly expressed in bone and cartilage containing tissue, as well as in osteogenic, preosteoblastic and chondroblastic cell lines. To determine if the human homolog mapped to 1q21 within the pyknodysosotosis interval, radiation hybrid mapping was performed using primers corresponding to the 3′ end of the human sequence. The human gene maps between markers WI-497 and D1S305, in the same location as the pycnodysostosis locus, validating ECM1 as a positional candidate for the disease. Shortly thereafter, we and others (Gelb et al. 1996, Johnson et al. 1996) identified the gene encoding cathepsin K, a cysteine protease that cleaves collagen and also maps to chromosome 1q21, as the pyknodsystosis gene.

The human genomic structure and coding sequence of ECM1 were characterized. The genomic sequence consists of ten exons and nine introns spanning approximately 5.6 kb of sequence (Table I). The coding sequence is 1.6 kb and has 79% overall identity to the mouse cDNA sequence. In the mouse, alternative splicing deletes an internal exon, resulting in the tissue specific expression of this transcript primarily in tail, front paw and skin (Bhalerao et al., 1995). This exon corresponds to human exon 7, which contains complete codon triplets. The removal of this exon by alternative splicing would not disrupt the reading frame. RNA isolated from human skin fibroblasts demonstrated alternative splicing of ECM1 exon 7 (data not shown). Alternative splicing of ECM1 exon 7 may result in functional diversity of the protein.

Analysis of the deduced amino acid sequence in comparison to the mouse sequence showed regions of similarity as well as regions of dissimilarity (Fig. 1). Overall, the human protein sequence is 69% identical to the mouse sequence, taking into consideration a cysteine-free region of only 45% identity between human and mouse (Fig. 1). The C-terminal cysteine-containing region is 75% identical between human and mouse, suggesting that this region is functionally significant.

In particular, analysis of the cysteine doublets and single residues in the mouse (Bhalerao et al., 1995) suggested a pattern similar to what is seen in the serum albumin family proteins (Brown, 1976; Yang et al., 1985) and the Endo 16 protein (Soltysik-Espanola et al., 1994). This region of the protein may be involved in binding ligands within the extracellular matrix and localizing them to cell surface receptors. Interestingly, 28 out of 29 of the cysteine residues are conserved from mouse to human, with the one exception being in the signal peptide where the mouse cysteine residue at amino acid 13 is changed to a tyrosine residue in the human sequence. The spacing within the cysteine containing region of the protein is conserved as well, with the exception of one amino acid insertion in the human sequence. This arrangement still retains the overall pattern of cysteine residues found in the serum albumin family of proteins (Brown, 1976; Yang et al., 1985), suggesting that this region of the protein and the cysteine residues are important for the biological function of this protein. However, the exact role of ECM1 in bone and cartilage remains to be determined.

Acknowledgments

We thank Christiane Robbins from the NHGRI Sequencing Core Facility for technical assistance, and Doctors Paul Bornstein, Iain McIntosh and Danilo Tagle for their helpful discussions. This work was supported by the Division of Intramural Research, NHGRI.

References

Bhalerao, J., Tylzanowski, P., Filie, J.D., Kozak, C.A. and Merregaert, J. : Molecular cloning, characterization, and genetic mapping of the cDNA coding for a novel secretory protein of mouse. *J. Biol. Chem. 270:* 16385–16394, 1995.

Brown, J.R.: Structural origins of mammalian albumin. *Fed. Proc. 35:* 2141–2144, 1976.

Gelb, B.D., Shi, G.-P., Chapman, H.A. and Desnick, R.J.: Pycnodysostosis, a lysosomal disease caused by cathepsin K deficiency. *Science 273:* 1236–1238, 1996.

Johnson, M.R., Polymeropoulos, M.H., Vos, H.L., Ortiz de Luna, R.I. and Francomano, C.A.: A nonsense mutation in the cathepsin K gene observed in a family with pycnodysostosis. *Gen. Res. 6:* 1050–1055, 1996.

Lennon, G., Auffray, C., Polymeropoulos, M. and Soares, M.B.: The I.M.A.G.E. Consortium: An integrated molecular analysis of genomes and their expression. Genomics 33: 151–152, 1996.

Shapiro, M.B. and Senapathy, P.: RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. 15: 7155–7175, 1987.

Soltysik-Espanola, M., Klinzing, D.C., Pfarr, K., Burke, R.D. and Ernst, S.G.: Endo16, a large multidomain protein found on the surface and ECM of endodermal cells during sea urchin gastrulation, binds calcium. Dev. Biol. 165: 73–85, 1994.

Walter, M., Spillett, D., Thomas, P., Weissenbach, J. and Goodfellow, P.: A method for constructing radiation hybrid maps of whole genomes. Nat. Genet. 7: 22–28, 1994.

Yang, F., Brune, J.L., Naylor, S.L., Cupples, R.L., Naberhaus, K.H. and Bowman, B.H.: Human group-specific component (Gc) is a member of the albumin family. Proc. Natl. Acad. Sci. USA 82: 7994–7998, 1985.

Dr. Douglas J. Wilkin, 10 Center Drive, MSC 1852, Building 10, Room 10C101, Bethesda MD 20892–1852.

Received June 26, 1997; accepted August 18, 1997