Genomic Organization and Chromosomal Localization of the Human Cathepsin L2 Gene

Reiko ITOH,1,2 Shoko KAWAMOTO,1 Wakako ADACHI,2 Shigeru KINOSHITA,2 and Kousaku OKUBO1,*

Institute for Molecular and Cellular Biology, Osaka University, 1-3 Yamada-Oka, Suita, Osaka, 565-08711 and Department of Ophthalmology, Kyoto, Prefectural University of Medicine, Kawaramachi, Hiroko-ji, Kamigyō-ku, Kyoto, 602-0841, Japan2

(Received 19 February 1999)

Abstract

Cathepsin L2 is a recently described cysteine proteinase with high sequence homology to cathepsin L and other members of the papain superfamily of cysteine proteinases. Its expression is regulated in a tissue-specific manner and is high in thymus, testis and cornea. In the present study, the entire gene sequence, including 5' and 3' flanking region, and chromosomal localization of human cathepsin L2 were determined. The gene spans approximately 6.4 kb and consists of eight exons and seven introns. Genomic organization was similar to human cathepsin L and more than 50% similarity was found between the first introns of cathepsin L and L2, suggesting that they diverged late in evolution. The transcription initiation site, determined by primer extension, was 198 nucleotides from the first ATG. The 5' flanking region lacks a TATA box but has one SP1 site. The gene was mapped to chromosome 9q21-22 by fluorescence in situ hybridization and the distance from cathepsin L was determined to be 15 cM by compiling radiation hybrid mapping results with a genetic map.

Communicated by Yoshiyuki Sakaki
* To whom correspondence should be addressed. Tel. +81-6-879-7991, Fax. +81-6-877-1922, E-mail: kousaku@imcb.osaka-u.ac.jp

Human cathepsin L2 (formerly called cathepsin V) is a recently described cysteine proteinase with no known ortholog.1,2 Based on its, primary structure and activity, cathepsin L2 is categorized as a member of the cysteine proteinases, which belongs to the papain superfamily.3 In this group, genes for cathepsins L, S, and K are thought to have evolved from a common ancestral gene prior to mammalian divergence because the sequence conservation between orthologs of different mammals are much higher than those among paralogs.3 Cathepsin L2 may have originated from ancestral cathepsin L much later because cathepsin L and cathepsin L2 have high amino acids identity (77%) and they are almost equally similar to cathepsin L in mouse (72% and 75%, respectively) and in other mammals.

Despite high structural similarities, cathepsins have discrete expression patterns. Cathepsins B, H and L have ubiquitous distributions,4-6 whereas S is highly expressed in spleen and lung7 and K is abundant in osteoclasts.8 The expression of isolated cathepsin L2 is more restricted. Significant expression of cathepsin L2 has been detected only in thymus, testis, and cornea among the more than 10 tissues tested.1,2 To assess the evolutionary relationship of the cathepsin L2 gene with other cathepsins and to provide a clue to understand its unique expression pattern, we have determined the entire gene sequence including 5' and 3' flanking regions and mapped it by FISH and RH panel.

A human genomic library in the Lambda FIX II vector (Stratagene, La Jolla, CA) was screened with a 32P-labeled full-length cathepsin L2 cDNA.1 One positive phage clone was isolated and used as a template for sequence analysis. A nucleotide sequence of 8813 bp, including the entire cathepsin L2 gene revealed that it consists of eight exons (Fig. la) ranging from 64 to 387 nt and seven introns ranging in size from 100 to 1673 nt, which spans approximately 6.4 kb (Fig. 1b). The exon-intron boundary sequences conformed to the GT/AG rule9 and coding sequence starts within exon 2. More than 50% similarity was found between the first introns of cathepsin L and L2, although it is interrupted by about a 300 bp Alu insertion in cathepsin L. Such homology was not observed between introns of human L or L2 and mouse cathepsin L.

The major transcription start site was determined by primer extension (Fig. 2a). There is neither a TATA nor a CCAAT box. Only one 9-bp SP1 site (GGCGCGCCC) was found 41 bp upstream of the major start site (P1). The GC content within this 681 bp was 51% (Fig. 2b). Of the human cathepsins in the papain proteinase family, cathepsins S, B, and K have been subjected to promoter sequence analysis. They all lack TATA boxes. The promoter of the cathepsin B gene has greater than 80% GC content with clusters of 15 SP1 sites, which is consis-
tent with its ubiquitous expression pattern. In contrast, cathepsins S and K have promoters with neutral GC content and few SP1 sites, similar to L2. Such featureless promoters seem to be characteristic of the cathepsin genes with tissue-regulated expression.

The cathepsin L2 gene was mapped by FISH to the same cytogenetic band (9q21-22) as cathepsin L4 (Fig. 3). The distance between them was estimated by typing the Gene Bridge 4, human/hamster radiation hybrid panel (Research Genetics, Inc., Huntsville, AL) using primers 5'-GAACTGACCAAACGCTTAT-3' and 5'-CAATGAGTCTTTGATATCAT-3', that specifically amplify the cathepsin L2 gene.1 The results submitted to the RH server at MIT (http://www-genome.wi.mit.edu) placed this gene between framework markers D9S197 and WI-6378, which corresponds to a region approximately 15 cM telomeric to cathepsin L in the genetic map. This contrasts with the finding that dissimilar paralogs cathepsins K and S lie within 150 kb on human chromosome 1q21.10 In this region, there is a short syntenic homology with mouse chromosome 13 where cathepsin L, FACC, SYK and GAS1 are contained in common (http://www.ncbi.nlm.nih.gov/Homology). In human, FACC, the nearest neighbor gene of cathepsin L, has been mapped11 between the same markers that flank cathepsin L2. In the mouse, FACC maps 6 cM telomeric to cathepsin L.12

In dbEST there are several mouse expressed sequence tags (ESTs) with high similarity to human cathepsin L2 that are not identical to mouse cathepsin L, including EST AA013726, which was used by another group to isolate human cathepsin L2.2 However, these ESTs are all more similar to cathepsin L than to L2, and probably correspond to the “cathepsin L-related mRNA” in rat (L14776) rather than the mouse ortholog of cathepsin L2.

In summary, cathepsin L2 has genomic organization that is highly similar to human cathepsin L. The sequence conservation even in introns supports the idea that they diverged late in evolution. Although these two homologs are mapped in the same cytogenetic band, they lie about 15 cM apart. The immediate upstream region is characterless, as in the case of other tissue-regulated cathepsins. Analysis of 5' flanking of cathepsin L gene may clarify the differential expression mechanism of these two highly related genes.

Acknowledgments: This work was supported by a
Figure 2. Identification of human cathepsin L2 transcription initiation sites and sequence of the upstream flanking region.
a. Twenty micrograms of human corneal epithelial cell total RNA\textsuperscript{13} was used for the primer extension (Pe) reaction with an end-labeled antisense primer that spans nucleotides +24 to +2 of the cDNA. The primer extension products were 148 and 131 bp in length, which corresponds to 198 nt (P1) and 181 nt (P2) from the first ATG, respectively. Conventional sequencing reaction ladder was used as a size marker. b. Nucleotide sequence of the 5' flanking region and exon 1 of the cathepsin L2 gene. Numbers in the margins are the nucleotide positions relative to the major transcriptional initiation site (P1). The putative SP1 binding site and an Alu repetitive element are underlined. The primer used in the primer extension reaction (AS) is also marked.

Figure 3. Fluorescence in situ hybridization of human cathepsin L2. Purified insert of a Lambda FIX clone containing the human cathepsin L2 gene was provided to Genome Systems (St. Louis, MO) for FISH mapping. Two pairs of signals on band 9q21-22 are indicated (arrows).
References