Gene organization of the small subunit of human calcium-activated neutral protease

Sanae Miyake, Yasufumi Emori* and Koichi Suzuki

Department of Molecular Biology, The Tokyo Metropolitan Institute of Medical Science, 3-18 Honkomagome, Bunkyo-ku, Tokyo 113, Japan

Received 1 September 1986; Revised and Accepted 20 October 1986

ABSTRACT
The gene for the small subunit of human calcium-activated neutral protease was isolated and sequenced. It is 11 kb long and comprises 11 exons. No TATA or CAT box was found upstream of the possible transcription initiation sites, but there are three so-called G-C box sequences and one G-C box-like sequence, which are usually found in "house-keeping" genes. The first exon (exon 1) contains only the 5'-noncoding sequence and exon 2 encodes the Gly-rich hydrophobic domain. Each of the four calcium-binding loop regions is encoded by one exon (exons 7-10). The intron breakpoints in the C-terminal calcium-binding domain (exons 4-11) completely coincide with those of the chicken large subunit gene. These findings suggest that the small and large subunits have evolved from the same ancestral calcium-binding protein and have retained the original gene organization.

INTRODUCTION
Calcium-activated neutral protease (CANP) is an intracellular thiol protease found in higher animals (1 - 5). The enzyme requires calcium ions for its activity and is probably involved in various cellular processes mediated by calcium. It is composed of a large subunit (80K; M.W. ca.80,000) and a small subunit (30K; M.W. ca.30,000). 80K is a catalytic subunit containing a protease domain and a calcium-binding domain (6). 30K contains a Gly-rich hydrophobic domain at the N-terminus which interacts with the cellular membrane (7) and a calcium-binding domain at the C-terminus (8, 9). The enzyme activity is regulated by two moles of calcium-binding domain in the two subunits. These two calcium-binding domains have homologous amino acid sequences (about 50% homology) and contain four consecutive E-F hand structures, which are typical calcium-binding sites found in calmodulin and troponin C. These two
domains can be regarded as "calmodulins" on the basis of the sequence homology and the distribution of the E-F hand structures along the peptide chain (5, 6, 9).

An E-F hand structure is composed of a loop region flanked by two α-helical regions and calcium binds to the loop region. Calcium-binding proteins with four E-F hand structures are thought to be the result of two steps of gene duplication from a gene coding one E-F hand structure (10, 11). In this respect, the gene structures of E-F hand proteins such as CANP and calmodulin are quite interesting as to their evolutionary processes.

The gene structure of chicken 80K has already been determined and the results for the calcium-binding domain are consistent with the gene duplication idea (12). The gene organization of the chicken calmodulin gene is different from that of "calmodulin" in chicken CANP 80K and cannot apparently be explained in terms of the above idea (13). Here, we determined the gene structure of human CANP 30K to clarify the evolutionary correlation between the two calcium-binding domains in CANP. The gene organizations of CANP and other calcium-binding proteins are discussed in terms of their evolutionary processes.

MATERIALS AND METHODS

Isolation and characterization of genomic clones

The human genomic library was kindly provided by Dr T. Maniatis (14). About 10^6 plaques were screened using the nick-translated cDNA fragment of the rabbit CANP 30K subunit (9) as a probe. DNA from the isolated phages was digested with EcoRI and then cloned into plasmid pUC9 (15). For DNA sequence analysis, the DNA inserts were further digested with appropriate restriction enzymes and then subcloned into convenient restriction sites of plasmids pUC9 and pUC18 (16). DNA sequencing was performed by the dideoxy chain termination procedure (17) or the chemical modification method of Maxam and Gilbert (18).

S1 nuclease mapping analysis

S1 nuclease analysis was carried out according to Maniatis et al. (19). The 1.5 kb EcoRI-SmaI fragment (see Figure 1) was subcloned into pUC18. This plasmid was digested with SmaI and
then end-labelled with T4 polynucleotide kinase and $\gamma^{32}\text{P}-d\text{ATP}$.

The EcoRI-SmaI fragment was recovered after digestion with EcoRI. The fragment was mixed with the total RNA of human spleen in 30μl of the hybridization buffer (80% formamide, 40mM Pipes-HCl, pH6.4, 0.4M NaCl, 1mM EDTA). Yeast tRNA and rabbit rRNA were also used as control RNAs. The solution was heated at 85 °C for 15 min and subsequently incubated at 63 °C overnight. After incubation, the solution was diluted with 300μl of S1 buffer (0.28M NaCl, 50mM acetate buffer, pH4.5, 4.5mM ZnCl$_2$, 1000 U/ml S1 nuclease) and then incubated at 45 °C for 30 min. The protected DNA fragments were analyzed on 7M urea-8% polyacrylamide gel.

Southern blot hybridization

Human genome DNA was isolated from lymphocytes and then digested with several restriction enzymes. The digested DNA (2μg each) was separated on 0.8% agarose gel and then transferred to a nitrocellulose filter according to Southern (20). Hybridization was carried out at 65 °C overnight in the presence of 0.2% SDS, 1× Denhardt’s solution (19), 1M NaCl, 50mM Tris.HCl, pH8.0, 10mM EDTA, 100μg calf thymus DNA and the nick-translated probe (1/10$^6$ cpm/ml). The filter was finally washed at 65 °C under stringent conditions (0.1× SSC, 0.1% SDS).

RESULTS

Isolation and sequencing of the gene for human CANP 30K

On screening of a human gene library (14) with rabbit CANP 30K cDNA as a probe (9), two types of clones containing 17 kb DNA in total were isolated. Inserts of the phage clones were subcloned into pUC9 or pUC18 and sequenced. Exons were identified by comparing the nucleotide sequence with the cDNA sequence (9).

The organization of the human CANP 30K gene is summarized in Figure 1, and the results of sequence analyses are shown in Figure 2. The gene is about 11 kb long and comprises 11 exons. Exon 1 contains only the 5'-noncoding sequence. The coding region starts at the 16th residue of exon 2. Exons 3-10 contain only the coding sequence. Exon 11 encodes the C-terminal eight amino acid residues and the total 3'-noncoding sequence. Exon 3
Fig. 1. Organization of the gene for the human CANP 30K subunit and the sequencing strategy. The regions corresponding to two independent clones, λ2 and λ4 are indicated at the top. Exons 1 to 11 are indicated by either hatched boxes (untranslated sequences) or solid boxes (protein-coding sequences), and open boxes denote introns. The restriction sites used for subcloning are shown above the boxes: EcoRI (E), HindIII (H), BamHI (B), PstI (P), SmaI (S), XhoI (X), BglII (Bg), HincII (Hc), PvuII (Pv), HaeIII (Ha) and Rsal (R). The arrows below the boxes indicate the directions and lengths of sequencing. The DNA fragment used for the S1 nuclease mapping analysis is indicated by ___ below the λ4 insert.

is the shortest (34bp) and exon 11 the longest (534bp). Most of the exons are shorter than 100 bp. All of the introns begin with the sequence, GT, and end with the sequence, AG, and other boundary sequences are consistent with the consensus sequence (21). The genomic DNA differed from the cDNA in four nucleotide sequences. As shown in Figure 2, one is located in the 5'-noncoding region and the other three in the 3'-noncoding region. These differences are probably due to nucleotide sequence polymorphism.

Transcription initiation sites

Comparison of the length of the cDNA for human 30K with that of mRNA by Northern blot analysis (22) indicated that nearly full-length cDNA had been cloned. However, no TATA or CAT box, general transcription initiation sequences of eukaryotes (23), existed upstream of the 5'-end of the cDNA. In order to determine the initiation site of transcription, S1 mapping analysis was performed. The 1.5 kb EcoRI-SmaI fragment containing the about 1.3 kb 5'-upstream region from the SmaI
Fig. 2. Nucleotide sequence of the human CANP 30K gene. The deduced amino acid sequence is indicated above and the nucleotide sequences different from the cDNA sequence are shown below the genomic sequence. Arrows indicate the possible initiation sites for transcription (see Figure 3). The 5' end of the longest cDNA is indicated by an asterisk (22). The G-C box sequences are double-boxed, and the G-C box-like sequence is single-boxed. The polyadenylation signal sequence (AATAAA) is underlined. The nucleotide residues are numbered beginning with the first base of the methionine codon.
Fig. 3. S1 mapping analysis of the 5'-end of the CANP 30K gene. The 1.5 kb EcoRI-SmaI fragment uniquely end-labelled at the SmaI site was hybridized with human spleen total RNA (lane 1), yeast tRNA (lane 2) or rabbit rRNA (lane 3). The protected DNA fragments that appeared only in lane 1 are indicated by arrows and numbered.

site in exon 1 (Figure 1) was used as a probe. As shown in Figure 3, many bands appeared even in control experiments with tRNA or rRNA. As this region is rich in G/C residues, this is presumably due to palindromic structures which are often formed in G/C rich regions. In spite of the high background, at least
Fig. 4. Southern blot analysis of the human CANP 30K gene. (A) Human genomic DNA digested with EcoRI (lane 1), HindIII (lane 2), PstI (lane 3) or BamHI (lane 4) was subjected to Southern hybridization analysis with the cDNA for human CANP 30K as a probe. The numbers on the left are the size markers (kb). (B) Restriction map of the human CANP 30K gene. The seven regions of the cDNA used as a probe are shown at the top. Only the lengths of the DNA fragments hybridized with the probe are shown.

six bands (arrows, Nos. 1-6, in Figure 3) appear to be specific for the human mRNA fraction (lane 1) and correspond to possible initiation sites. The shorter RNA bands (Nos. 1 and 2) are much more intense than the longer RNA bands (Nos. 3-6). There are three G-C boxes (GGCGGG) and one G-C box like sequence (CCCGCC) upstream of the transcription initiation sites (Figure 2, double-boxed and single-boxed, respectively). Some of eukaryotic genes are known to have G-C box sequences in place of TATA and CAT boxes as the transcription initiation signal and
multiple transcription initiation sites (24 - 27). The results obtained suggest that these six sites are probably the transcription initiation sites of the mRNA.

The 5'-end of the cDNA, indicated by an asterisk in Figure 2, is, however, in the upstream region of the four possible sites (Nos. 1-4). The longest cDNA clone may correspond to the mRNA initiated at site No.5 or 6, though the possibility that these bands are artifacts cannot be excluded.

Identification of the gene for human CANP 30K

To determine the number of the gene, a total genomic Southern experiment was performed. Human total DNA was isolated from lymphocytes and digested with EcoRI (Figure 4A, lane 1), HindIII (lane 2), PstI (lane 3) or BamHI (lane 4). Human 30K cDNA (22), shown at the top of Figure 4B, was used as a probe and hybridized with these restriction fragments. The pattern obtained is essentially consistent with the restriction maps of the isolated clones (Figure 1 and 4). Clear bands are 8.4 kb and 5.0 kb bands (lane 1), 8.2 kb, 6.2 kb and 2.6 kb bands (lane 2), 3.6 kb, 1.4 kb and 0.6 kb bands (lane 3), and doublet 6.0 kb bands (lane 4). Besides these major bands, 2.8 kb and 1.4 kb bands (lane 1), and a few faint longer bands (lanes 2-4) also appeared. These bands might be due to EcoRI* activity (lane 1), incomplete digestion (lanes 2-4) or the presence of another homologous sequence in the genome DNA (lanes 1-4), although these possibilities should be further investigated. From these results, the gene is considered to be essentially a single copy gene.

Comparison with the chicken CANP 80K gene

Previous studies revealed that the C-terminal regions of 80K and 30K have homologous amino acid sequences and contain four calcium-binding sites (8, 9). To elucidate their evolutionary relationship, the intron positions of the 30K gene were compared with those of the 80K gene (12). In these regions, the nucleotide and amino acid sequences of human CANP 30K and chicken 80K can be aligned as shown in Figure 5. The intron positions for the 80K and 30K genes shown by upward or downward arrows, respectively, are located exactly at the same positions in terms of the amino acid sequence and the nucleotide sequence.
Fig. 5. Comparison of the nucleotide and amino acid sequences and intron positions of the two subunits of CANP. Only the sequences of the calcium-binding domains corresponding to the C-terminal regions of both subunits are shown and aligned. The asterisks indicate identical nucleotide residues between the chicken 80K and human 30K. The common amino acid residues are shown by boldface letters. The loop regions of the E-F hand structures are shadowed and numbered. The intron positions of the 30K are indicated by downward arrows and capital letters, and those of the 80K by upward arrows and small letters.

as well. This clearly indicates that these regions of the two subunits are derived from a common ancestral gene and also suggests that both genes retain the original structures present at the time of their divergence. Furthermore, it is clear that each of the four calcium-binding loops in both subunits, shadowed in the Figure, is encoded by one exon.

Although the positions of introns are conserved, the sizes of the introns are different for these two genes; those of 30K are 0.167 kb (A), 0.95 kb (B), ca.2.4 kb (C), 0.1 kb (D), 0.14 kb (E), ca. 3 kb (F) and 0.174 kb (G), while those of 80K are ca.0.2 kb (a), 0.15 kb (b), 0.9 kb (c), 0.1 kb (d), 0.5 kb (e), 0.1 kb (f) and 0.2 kb (g) (12). During evolution, all of the
intron positions were retained but the lengths of the introns changed.

**DISCUSSION**

Introns are thought to mediate the gene duplication of one domain (28, 29). The junction between the duplicated domains contains the intron and the introns show the gene history. The genes for calcium-binding proteins which have four calcium-binding sites composed of E-F hand structures are thought to be the result of the two steps of gene duplication of a gene encoding a peptide having only one calcium-binding site (10, 11). Each CANP subunit has four E-F hand structures and each calcium-binding loop is encoded by one exon (12). Thus, the genes for both subunits of CANP are essentially consistent with the gene duplication idea (Figure 5). As for other calcium-binding proteins, however, the gene structures of chicken (13) and rat calmodulins (30) are not consistent with the idea. Namely, some of intron positions are located in the middle of the calcium-binding loops. Presumably, in the case of calmodulin, gene rearrangement might occur soon after the gene duplication. Thus the genes may not retain the original gene structures.

Comparison of the positions of introns in various genes is one means of determining the evolutionary correlation. The cytokeratin (31), keratin (32) and protein C (33) genes were studied in this respect. In these cases, a close evolutionary relationship was detected on examining the intron positions of various proteins of a certain family, subfamily or superfamily. In the case of CANP, the intron breakpoints of 30K completely coincide with those of 80K, which indicates that the calcium-binding regions in the two subunits might have evolved from the same ancestral gene. Fusion of this ancestral gene with a thiol protease gene should have produced the 80K subunit gene, and fusion of the latter with a gene encoding a Gly-rich hydrophobic peptide should have given the 30K subunit gene. At least as far as the calcium-binding domain is concerned, the gene structures have been conserved and the original structures retained.
Since the results of S1 mapping of the promoter region of the CANP 30K gene were not sufficiently clear, primer extension analysis was further performed with two synthetic oligonucleotides complementary to positions -53 to -22 and -128 to -107 (Figure 2). Again, clear results were not obtained (data not shown). This region probably tends to form complicated secondary structures because of its high G/C content and strand elongation was inhibited.

The 5'-upstream region of the human CANP 30K gene has several features in common with house-keeping genes, such as the adenosine deaminase (24), hydroxy-3-methylglutaryl coenzyme A reductase (25) and hypoxanthine phosphoribosyltransferase (26, 27) genes. Common features are 1) the region is rich in G/C, 2) three G-C boxes and one G-C like box are present and 3) multiple transcription initiation sites exist. On the other hand, CANP activity is detected widely in various cells and tissues of higher animals. Moreover, CANP genes are expressed in all tissues so far examined, though the content of mRNA is low (ca.0.1% of total mRNA) (9, 33, 34). From these results the CANP gene can be regarded as a house-keeping gene and the G/C rich region will be the promoter region.

Further studies on the gene structures of other calcium-binding proteins, CANP-related proteins and house-keeping proteins will reveal the evolutionary and functional characteristics of these genes in more detail.

ACKNOWLEDGEMENTS

We wish to thank Dr T. Maniatis (Harvard University) for providing the human genomic DNA library. This work was supported in part by research grants from the Ministry of Education, Science and Culture, a Grant-in-Aid for New Drug Development, and grants from the National Center for Nervous, Mental and Muscular Disorders of the Ministry of Health and Welfare of Japan.

*To whom correspondence should be addressed
REFERENCES