cDNA cloning and sequencing of component C5 of proteasomes from rat hepatoma cells

Tomohiro Tamura[1], Keiji Tanaka[1], Atsushi Kumatori[1], Fumi Yamada[1], Chizuko Tsurumi[1], Tsutomu Fujiwara[2], Akira Ichihara[1], Fuminori Tokunaga[3], Rie Aruga[4] and Sadaaki Iwanaga[4]

[1]*Institute for Enzyme Research, The University of Tokushima, Tokushima 770,* [2]*Otsuka Pharmaceutical Company Ltd., Tokushima 771-01,* [3]*Department of Molecular Biology, Graduate School of Medical Science* and [4]*Department of Biology, Faculty of Science, Kyushu University 33, Fukuoka 812, Japan*

Received 26 February 1990; revised version received 16 March 1990

Proteasomes are multicatalytic proteinase complexes consisting of a set of non-identical polypeptide subunits. A cDNA for component C5 of rat proteasomes was isolated by screening a Reuber H4TG hepatoma cell cDNA library using synthetic oligodeoxynucleotide probes corresponding to partial amino acid sequences of the protein. The polypeptide deduced from the open reading frame consisted of 240 amino acid residues with a calculated molecular weight of 26479. Computer analysis revealed little similarity of C5 to other proteins reported so far. The primary structure of C5 showed partial sequence homology to that of another component C3, but no regions of homology with the sequence of component C2. Thus C5 is concluded to be a new type of subunit of the proteasome complex.

cDNA cloning; Multicatalytic proteinase; Proteasome; Subunit C5

1. INTRODUCTION

Proteasomes are multisubunit complexes with a molecular mass of approximately 750 kDa [1–3]. The reason why proteasomes have such an unusual, complex structure is unknown, but the individual subunits may each have different functions. For determining the functions of these proteasomal multisubunit complexes, it is essential to obtain information about the structure–function relationship of individual subunits. Accordingly, we are attempting to determine the primary structures of all the subunits of proteasomes by recombinant DNA techniques. Recently, we have isolated and sequenced cDNAs for two components, C2 [4] and C3 [5], of rat liver proteasomes and have found that these are the products of independent genes and are novel proteins, differing from any other known protein. In this paper, we report the cloning and sequencing of a full-length cDNA for another component, C5, of rat proteasomes.

2. MATERIALS AND METHODS

Proteasomes were purified to homogeneity from rat liver by conventional chromatographic techniques, as previously reported [2].

Correspondence address: K. Tanaka, Institute for Enzyme Research, The University of Tokushima, Tokushima 770, Japan

Abbreviations: SDS-PAGE, sodium dodecyl sulfate-polyacrylamide gel electrophoresis; HPLC, high-performance liquid chromatography; bp, base pairs

Component C5 was isolated from the purified proteasomes by reversed-phase HPLC on a Cosmosil $5C_4$-300 column [3]. The C5 protein was reduced, *S*-pyridylethylated, and then digested with lysyl endopeptidase. The amino acid sequences of the protein and peptides were determined with a gas-phase sequencer (Applied Biosystems, model 477A), and 120A phenylthiohydantoin analyzer on-line system [4].

A cDNA library of Reuber H4TG hepatoma cells was constructed in a phagemid expression vector, Bluescript KS^+ (Stratagene) as described previously [5]. For isolation of cDNAs for component C5, about 30000 transformants were screened by hybridization with oligodeoxyribonucleotide probes that had been synthesized in an Applied Biosystems Model 380B DNA Synthesizer and labeled at their 5′-end with [γ-^{32}P]ATP and T4 polynucleotide kinase. Colony hybridization was carried out by a previously reported method [4]. DNA sequencing was carried out by the dideoxy chain-termination method with a 7-DEAZA Sequencing Kit (Takara Shuzo) and Sequenase (Toyobo).

RNA hybridization analysis was performed as described before [4] using 10 μg of poly(A)$^+$ mRNA extracted from various tissues of rats and a probe (700 bp *Xho*I-*Sau*3AI fragment of C5 cDNA) labeled by the multiprime DNA labeling method.

3. RESULTS AND DISCUSSION

3.1. *Protein sequencing of component C5*

Previously, we reported the separation of multiple components of proteasomes from rat liver by reversed-phase HPLC [3]. First, 10 major components were separated on a Cosmosil $5C_4$-300 column and named component 1 (C1) to component 10 (C10) in order of their elution. Component C5 was recovered at a 50% acetonitrile concentration. Fig. 1 shows the profiles on SDS-PAGE of all the components of rat liver proteasomes and the isolated component C5. On SDS-

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/90/$3.50

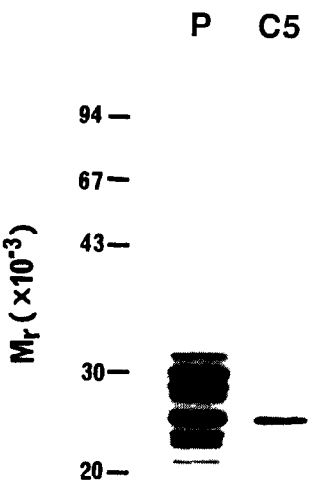


Fig. 1. Profiles of rat liver proteasomes and isolated component C5 on SDS-PAGE. Proteasomes (25 μg) and component C5 (approximately 2 μg) were subjected to SDS-PAGE. M_r, molecular weight markers; P, purified proteasome complexes; C5, isolated C5. Proteins were stained with Coomassie brilliant blue.

PAGE analysis, the molecular weight of C5 was determined to be 25100 ± 900 [3,4].

On automated Edman degradation analysis, the N-terminal amino acid residue of the purified C5 was determined to be Arg and its N-terminal 20 amino acids were sequenced (broken line in Fig. 2). To obtain further information on the primary structure of its internal region, samples of C5 were reduced, *S*-pyridylethylated and digested with lysyl endopeptidase. The resulting fragments were resolved by reversed-phase HPLC on a Chemcosorb 7-ODS-H column (data not shown) and their partial primary structures (approximately 90 amino acid residues) were analyzed by Edman degradation (solid lines, in Fig. 2).

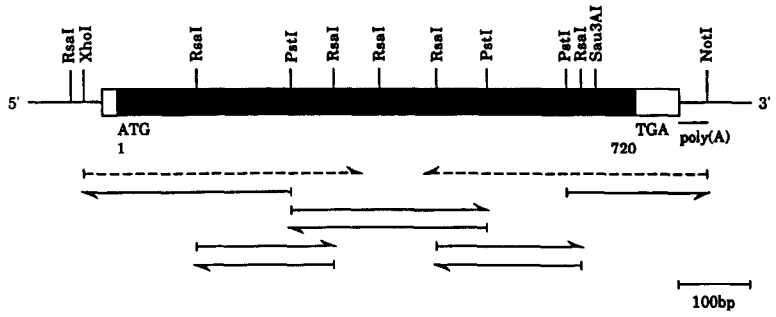
3.2. *Isolation of cDNA clones encoding component C5*

Recently we found that the mRNA levels of components C2 and C3 in Reuber H4TG hepatoma cells were approximately 10 times those in normal rat liver (unpublished data). To isolate proteasome cDNA clones more efficiently, we constructed a new cDNA library with the phagemid Bluescript KS^+ vector using poly(A)$^+$ RNAs extracted from the H4TG cells. We screened this library by hybridization with synthetic deoxyribonucleotides as probes. These probes were synthesized as follows. Of the C5 fragments obtained by proteolytic digestion, parts of the sequences of three fragments (Lys-Asn-Met-Gln-Asn-Val, Asp-Asn-Gln-Val-Gly and Tyr-Ala-Phe-Asn-Gly) with minimum codon ambiguity were selected and the oligonucleotide probes 5′-ACRTTYTGCATRTTYTT-3′, 5′-CCNACYTGRTTRTC-3′ and 5′-CCRTTRAANGCRTA-3′ (N = AGTC, Y = TC, R = AG) were synthesized. We first screened about 30000 colonies of the H4TG cell cDNA library with 16 mixtures of 17mer oligonucleotides. For second screening, we used two sets of 32 mixtures of each of these 14mer oligonucleotides. Eleven cDNA clones that gave strongly positive signals with both probes were isolated from the library by colony hybridization techniques. Of these, the clone (about 800 bp in length including a poly(A) tail) that carried the largest cDNA insert was subjected to cDNA sequencing.

3.3. *Primary structure deduced from the cDNA sequence*

The nucleotide sequence of the C5 cDNA clone was determined by the strategy illustrated in Fig. 2 (upper panel). The resulting nucleotide sequence of the cDNA and the primary structure of the C5 deduced from the cDNA sequence are shown in Fig. 2 (lower panel). The sequence of 799 nucleotides included the entire coding region and 5′- and 3′-non-coding regions. The 3′-non-coding region consisted of 61 nucleotides, excluding the poly(A) tail. A putative polyadenylation signal (ATTAAA), which is common to eukaryotic mRNAs, was located 20 nucleotides upstream from the poly(A) addition site. Part of the poly(A) sequence (approximately 14 nucleotides) is seen at the 3′-end of this cDNA sequence. The longest open reading frame (ORF) is 720 nucleotides and corresponds to a protein of 240 amino acids. From this deduced sequence, the molecular weight of C5 was calculated to be 26479. We concluded that ATG, located at nucleotides 1–3, is the initiation codon, because the longest ORF is usually thought to be the coding sequence of a protein and because this ATG is surrounded by sequences that are close to the consensus sequence for translation initiation [6]. However, two other possible ATG initiation codons are present downstream of the predicted initiation codon ^{1}ATG^{3} for C5 at nucleotides 46–48 and 79–81 (boxed in Fig. 2). Particularly, as the amino acid sequence of the N-terminal region of C5 determined by Edman degradation corresponded to nucleotides 82–141, the possibility that ^{79}ATG^{81}, just before the N-terminal Arg residue determined chemically, is the initiation codon for C5 cannot be ruled out. However, the N-termini of all other proteasome subunits except C5 were blocked [4,5] and the nucleotide sequences surrounding the ^{79}ATG^{81} clearly differed from the conserved consensus sequence for translation initiation [6], and so the N-terminal Arg residue of C5 determined chemically may be produced by limited proteolysis during purification of the protein. Further studies are required to determine the exact N-terminus for C5.

The amino acid sequence shown in Fig. 2 is actually that of C5 of proteasomes, because the partial sequences of several fragments (approximately 50% of the total residues) determined by Edman degradation




```
                                                                5'-CAGCAGGAGTTAGGCGCG      -1

ATG CTT TCC ACT GCC GCT TAC CGG GAC CCG GAC CGA GAG CTG GTA ATG GGG CCT CAG GGC      60
Met Leu Ser Thr Ala Ala Tyr Arg Asp Pro Asp Arg Glu Leu Val Met Gly Pro Gln Gly      20

TCC GCC GGC CCT GTG CAG ATG CGT TTC TCG CCT TAT GCC TTC AAC GGA GGT ACT GTA TTG     120
Ser Ala Gly Pro Val Gln Met Arg Phe Ser Pro Tyr Ala Phe Asn Gly Gly Thr Val Leu      40

GCA ATT GCT GGA GAA GAT TTT TCC ATC GTT GCT TCA GAC ACT CGG TTG AGT GAA GGG TTT     180
Ala Ile Ala Gly Glu Asp Phe Ser Ile Val Ala Ser Asp Thr Arg Leu Ser Glu Gly Phe      60

TCA ATT CAT ACC CGA GAT AGC CCC AAA TGT TAC AAA CTA ACA GAC AAA ACA GTA ATT GGC     240
Ser Ile His Thr Arg Asp Ser Pro Lys Cys Tyr Lys Leu Thr Asp Lys Thr Val Ile Gly      80

TGC AGT GGC TTT CAT GGA GAT TGT CTC ACC TTG ACA AAG ATT ATT GAA GCA AGA TTA AAG     300
Cys Ser Gly Phe His Gly Asp Cys Leu Thr Leu Thr Lys Ile Ile Glu Ala Arg Leu Lys     100

ATG TAC AAG CAT TCC AAT AAC AAG GCC ATG ACA ACA GGG GCG ATT GCT GCA ATG CTG TCT     360
Met Tyr Lys His Ser Asn Asn Lys Ala Met Thr Thr Gly Ala Ile Ala Ala Met Leu Ser     120

ACC ATC CTG TAC TCA CGG CGT TTG TTC CCT TAC TAT GTC TAC AAC ATC ATT GAG GGA CTT     420
Thr Ile Leu Tyr Ser Arg Arg Phe Phe Pro Tyr Tyr Val Tyr Asn Ile Ile Glu Gly Leu     140

GAT GAA GAA GGA AAG GGA GCC GTG TAC AGC TTT GAC CCA GTA GGC TCT TAC CAG AGA GAC     480
Asp Glu Glu Gly Lys Gly Ala Val Tyr Ser Phe Asp Pro Val Gly Ser Tyr Gln Arg Asp     160
                                  K1
TCT TTC AAA GCC GGA GGC TCA GCA AGT GCC ATG CTG CAG CCG CTG CTC GAC AAC CAG GTT     540
Ser Phe Lys Ala Gly Gly Ser Ala Ser Ala Met Leu Gln Pro Leu Leu Asp Asn Gln Val     180
                     K2
GGC TTC AAA AAT ATG CAG AAT GTG GAG CAT GTC CCC CTG ACT CTG GAC AGA GCC ATG AGG     600
Gly Phe Lys Asn Met Gln Asn Val Glu His Val Pro Leu Thr Leu Asp Arg Ala Met Arg     200
                        K3
CTG GTG AAA GAC GTC TTC ATT TCT GCA GCC GAG AGG GAT GTG TAC ACT GGA GAC GCC CTC     660
Leu Val Lys Asp Val Phe Ile Ser Ala Ala Glu Arg Asp Val Tyr Thr Gly Asp Ala Leu     220
                        K4
AGG ATC TGC ATC GTG ACC AAA GAG GGC ATC AGG GAG GAG ACT GTT CCC CTT CGG AAA GAC     720
Arg Ile Cys Ile Val Thr Lys Glu Gly Ile Arg Glu Glu Thr Val Pro Leu Arg Lys Asp     240
                                 K5
TGATGTGGGGCCAGGGTTTGGAGCTGTGGCACGTTTGTTTTATTAAAAAGAAATGTAAAGG-3'                  781
===                                                                  ==========
```

Fig. 2. Structure of cDNA for component C5 of rat proteasomes. (Upper panel) Restriction endonuclease map of cloned cDNA for component C5 and sequencing strategy. The solid and open boxes show the coding region and 5'- and 3'-non-coding regions, respectively. Continuous lines indicate the sequence of the vector, Bluescript KS$^+$. The numbers below the boxes indicate the nucleotide numbers of the first nucleotide of the initiation codon, ATG, and the nucleotide before the termination codon, TGA. Sequenced regions are shown by horizontal arrows. The sequences of the 5'- and 3'-ends of cDNA (dashed arrows) were determined by direct sequencing of an isolated clone, using primers of the T3 or T7 promoter of Bluescript KS$^+$. Some fragments were obtained by subcloning after cleavages with *Rsa*I and *Pst*I and were sequenced with T3 and T7 primers (solid arrows). The bar represents 100 bp (base pairs). (Lower panel) Nucleotide sequence of the cDNA encoding component C5 and the amino acid sequence deduced from its longest open reading frame. Nucleotides are numbered in the 5' to 3' direction, beginning with the first residue of the putative initiation methionine codon ^{1}ATG^{3}. Two other possible translation initiation sites ^{46}ATG^{48} and ^{79}ATG^{81} are boxed. The nucleotides on the 5'-side are indicated by negative numbers. The predicted amino acid sequence of C5 is shown below the nucleotide sequence. Amino acid residues are numbered from the N-terminus. Broken and continuous lines show the amino acid sequences corresponding to those obtained by Edman degradation of purified component C5 and their fragments cleaved with lysyl endopeptidase, respectively. Amino acids shown by dotted lines in the sequences of K1, K2, and K4 were not identified by chemical analysis. The termination codon TGA is doubly underlined with broken lines. The possible polyadenylation signal (ATTAAA) is doubly underlined with continuous lines.

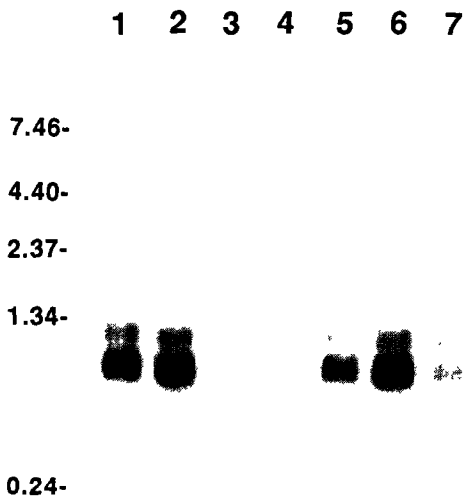


Fig. 3. RNA blot hybridization of rat liver C5 cDNA with poly(A)$^+$ RNAs from various rat tissues. Poly(A)$^+$ RNAs (10 μg) were used for Northern blot analysis. The *Xho*I-*Sau*3AI fragment (700 bp) of cloned C5 cDNA was used as a probe (Fig. 2). Rat liver (lane 1), kidney (lane 2), skeletal muscle (lane 3), heart (lane 4), lung (lane 5), spleen (lane 6) and brain (lane 7) were examined. The positions of RNA standards (from Bethesda Research Laboratories) are shown (in kilobases) on the left.

were found to be in excellent accordance with those deduced from the nucleotide sequence of cDNA (solid and broken lines in Fig. 2).

We searched for structural homologies of the nucleotide and amino acid sequences of C5 with those of other proteins using the data bases EMBL/GDB (European Molecular Biology Laboratory) and NBRF/PDB (National Biomedical Research Foundation). No proteins with overall similarity to C5 were found, suggesting that component C5 is distinct from other known proteins. The amino acid sequence of C5 partially resembled that of C3, but did not show any overall similarity to that of C2 or a 35 kDa component of *Drosophila* proteasomes reported by Haass et al. [7], although rat proteasome components C2, C3 and the *Drosophila* 35 kDa protein reveal marked sequence homology with each other. Thus C5 is thought to be a new type of proteasome component.

3.4. *Expression of the gene for component C5 in various rat tissues*

To investigate the tissue specificity of C5 gene expression, we examined the levels of C5-specific mRNA in various rat tissues. On Northern blot analysis, the mRNAs extracted from various rat tissues gave two hybridization bands of about 850 bp and 1150 bp (Fig. 3). At present, it is unclear whether both mRNAs encode C5 or whether the larger faint band is that of another protein homologous to C5. The levels of mRNA for component C5 varied significantly in the different tissues examined, but were very similar to those of the mRNAs for components C2 [4] and C3 [5], suggesting that proteasome complexes, presumably all the components, are expressed similarly in different cells and tissues.

Acknowledgements: This work was supported in part by a Grant-in-Aid for Scientific Research from the Ministry of Education, Science and Culture of Japan. We thank S. Kajiyama and Y. Nishina for performing amino acid analyses.

REFERENCES

[1] Rivett, A.J. (1989) Arch. Biochem. Biophys. 268, 1–8.
[2] Tanaka, K., Yoshimura, T., Kumatori, A., Ichihara, A., Ikai, A., Nishigai, M., Kameyama, K. and Takagi, T. (1988) J. Biol. Chem. 263, 16209–16217.
[3] Tanaka, K., Yoshimura, T., Ichihara, A., Ikai, A., Nishigai, M., Morimoto, M., Sato, M., Tanaka, N., Katsube, Y., Kameyama, K. and Takagi, T. (1988) J. Mol. Biol. 203, 985–996.
[4] Fujiwara, T., Tanaka, K., Kumatori, A., Shin, S., Yoshimura, T., Ichihara, A., Tokunaga, F., Aruga, R., Iwanaga, S., Kakizuka, A. and Nakanishi, S. (1989) Biochemistry 28, 7332–7340.
[5] Tanaka, K., Fujiwara, T., Kumatori, A., Shin, S., Yoshimura, T., Ichihara, A., Tokunaga, F., Aruga, R., Iwanaga, S., Kakizuka, A. and Nakanishi, S. (1990) Biochemistry, in press.
[6] Kozak, M. (1984) Nucleic Acids Res. 12, 857–872.
[7] Haass, C., Pesold-Hurt, B., Multhaup, G., Beyreuther, K. and Kloetzel, P.-M. (1989) EMBO J. 8, 2373–2379.