Volume 221, number 1, 43-47 FEB 05047 August 1987

# Nucleotide sequence of a rice glutelin gene

Fumio Takaiwa, Hiroyasu Ebinuma*, Shoshi Kikuchi and Kiyoharu Oono

*Department of Cell Biology, National Institute of Agrobiological Resources, Yatabe, Ibaraki 305, Japan*

Received 2 June 1987

A gene encoding the major rice storage protein, glutelin, was isolated and its nucleotide sequence was determined. The cloned gene (λINE 3) contains three short introns and codes for a prepro-glutelin protein of 499 amino acids ($M_r$ 56305) identical with that deduced from the type II glutelin cDNA. In the 5'-flanking region, a sequence similar to the legumin box characteristic to leguminous 11 S globulin genes can be detected around 60 bp upstream from the transcription start site.

Storage protein; Glutelin gene; $S_1$ nuclease mapping; *(Oriza sativa)*

## 1. INTRODUCTION

Glutelin is the major seed protein of rice, which accounts for approx. 80% of the total endosperm protein. It is encoded by a small multigene family (about 5 copies per haploid genome) [1]. Its expression is under strictly tissue- and developmental specific control. In order to elucidate this specific gene regulation, it is essential to isolate and characterize a genomic glutelin gene. We have previously isolated and determined the nucleotide sequences of glutelin cDNAs coding for full-length glutelin precursor [1,2]. Here, a rice glutelin gene was cloned and the structure was determined by sequencing a genomic fragment of 3.4 kb containing the entire RNA coding region as well as 900 bp of the 5'- and about 500 bp of the 3'-flanking regions. When compared with the cDNA sequences, it is found that the isolated glutelin gene is interrupted by three short introns and is functionally active in vivo. This is the first genomic nucleotide sequence of the rice storage protein gene to be described.

Correspondence address: F. Takaiwa, Department of Cell Biology, National Institute of Agrobiological Resources, Yatabe, Ibaraki 305, Japan

* Present address: Laboratory of Sericultural Science, Faculty of Agriculture, University of Tokyo, Tokyo, Japan

The nucleotide sequence presented has been submitted to the EMBL-GenBank database under the accession number Y00687

## 2. MATERIALS AND METHODS

Genomic DNA was prepared from dry rice embryos (*Oriza sativa* cv. Mangetsumochi) as described [1]. It was completely digested with *Bam*HI and fractionated on a sucrose gradient. A 12 kb DNA fragment containing the glutelin gene was recovered and then ligated to the *Bam*Hi arms of the phage EMBL 3. The concatemeric DNA was packaged in vitro. The phage plaques were screened by in situ hybridization [3] using a $^{32}$P-labeled pREE61 cDNA insert as a probe. DNA sequencing was carried out using the chemical [4] and dideoxy methods [5]. $S_1$ protection mapping was carried out as in [6]. The DNA sequence was analyzed by using tht GENETYX program (Software Development, Tokyo).

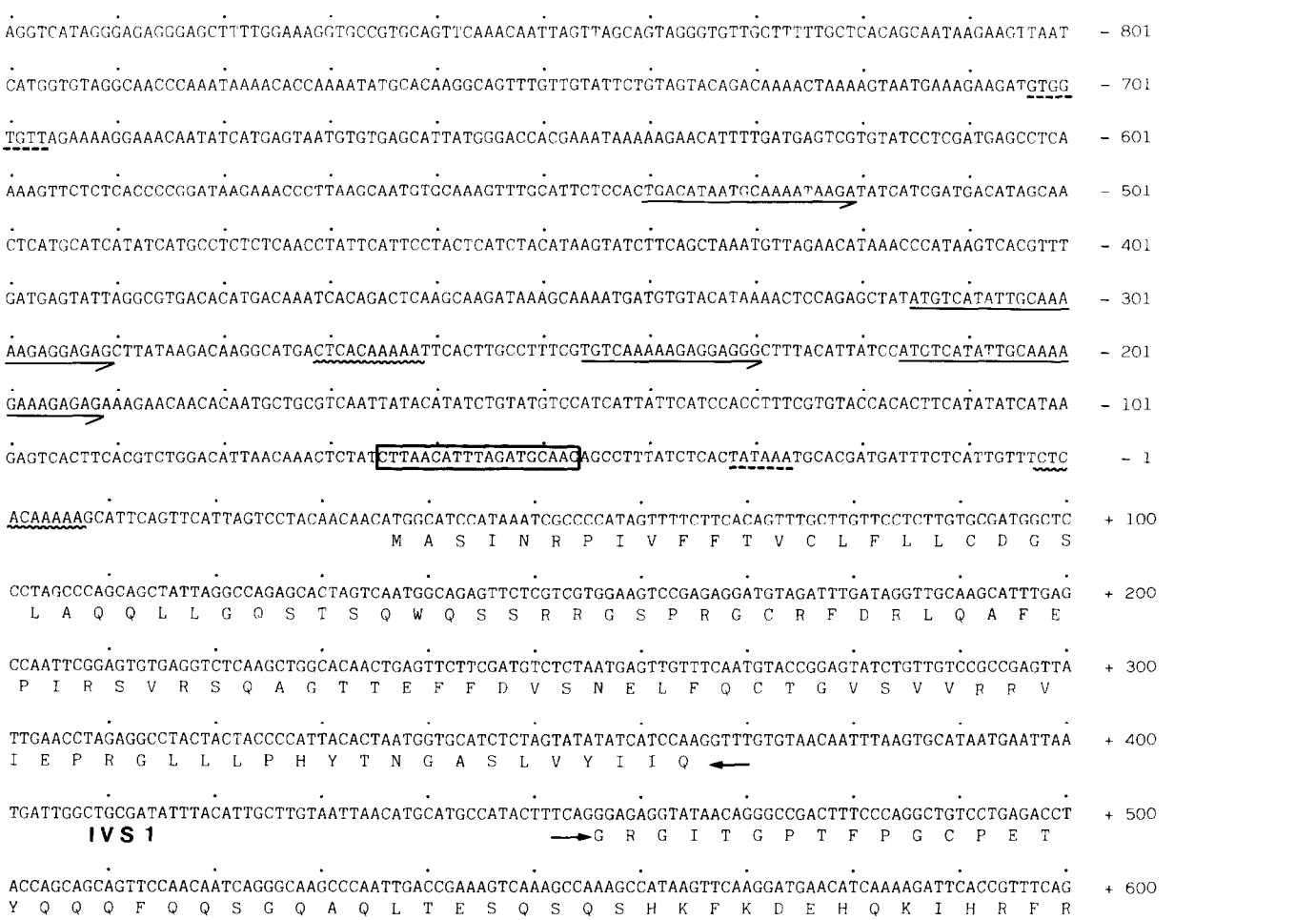
## 3. RESULTS AND DISCUSSION

The glutelin gene was screened by plaque hybridization using the pREE61 cDNA insert,

*Published by Elsevier Science Publishers B.V. (Biomedical Division)*
00145793/87/$3.50 

which resulted in the isolation of five positive clones. One of the clones, λINE 3, was chosen for further characterization as it strongly hybridized to the pRE61 cDNA. When this DNA was digested with several restriction enzymes and blot-hybridized with glutelin cDNA, it was shown that the glutelin gene was located in the middle region of the insert. Furthermore, it was also shown that the glutelin gene covering the whole of the transcribed region and the 5′- and 3′-flanking regions was confined to the 4.2 kb *Eco*RI fragment. This fragment was recloned into pUC118 to generate the subclone pREE771 and its insert was sequenced.

When the genomic sequence was compared with the previously determined cDNA sequences, the presence of introns was apparent (fig.1). This is the first such case to be found in the cereal seed storage protein genes. They occur between positions 367 and 455, 731 and 833, and 1314 and 1396. Therefore, the former two introns are located in the acidic subunit coding region and the latter in the basic subunit coding region. They range in size from 83 to 103 bp. Such a short length of introns is characteristic of plant genomic genes. The exon-intron junction sequences obey the AG/GT rule and show additional homology to the donor and acceptor conserved sequences derived from plant genes [7]. These introns are higher in A and T content than the coding region (intron 1, 68.5%; intron 2, 73.7%; intron 3, 67.4% vs coding region, 54.0%). It is known that leguminous 11 S globulin genes have two or three introns [8,9]. They are found at exactly the same positions as those of the rice glutelin gene. This fact reinforces the hypothesis that rice glutelin and leguminous 11 S



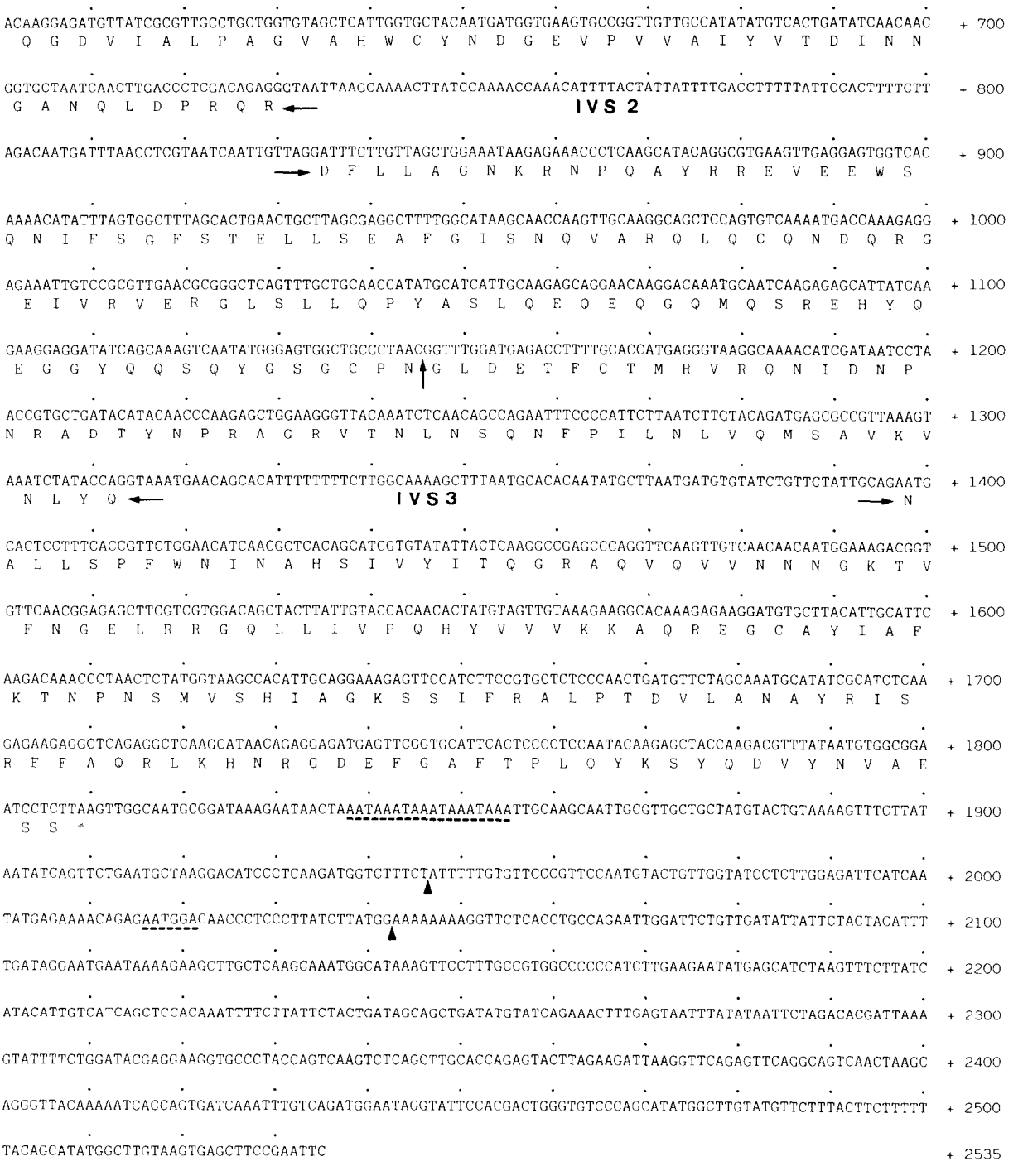



Fig.1. Nucleotide sequence (noncoding strand) and deduced amino acid sequence of a glutelin gene. Numbers are bp relative to the transcription start site. Consensus sequences (TATA box, enhancer core element and polyadenylation signals) are indicated by dashed lines. The legumin-box-like sequence is boxed. Direct repeat and CA-rich sequences are indicated by horizontal arrows and wavy lines. Arrowheads and vertical arrow indicate the poly(A) addition sites and the posttranslational processing site between the acidic and basic subunits.

globulin genes originate from a common ancestral gene, which has been suggested from the amino acid sequence homology between them [1].

The genomic glutelin gene contains a 1497 bp open reading frame coding for a 499-amino-acid polypeptide ($M_r$ 56305), which is the same size as that deduced from the type II glutelin cDNA [1]. This indicates that the isolated genomic clone (λINE 3) corresponds to the gene encoding type II glutelin. Therefore, it is not a pseudogene but an active gene, which is strongly transcribed in developing endosperm tissue.

The transcription start site was established to be A at position 1 by $S_1$ protection mapping (fig.2), which is 35 bp upstream from the translation start codon. The sequence surrounding this start site is TCAC, which is similar to the consensus sequence proposed by Dunsmuir [10]. It is noteworthy that a 10 bp direct repeat sequence identical with the transcription start sequence can be detected between positions −262 and −271. In the 5′-flanking region, the putative TATA box occurs 27 bp upstream from the transcription start site. There is, however, no apparent CAAT box at the expected position. A sequence homologous to the core enhancer element, GTGGTGTT, is observed 697 bp upstream from the transcription start site. An unusual feature of the 5′-flanking region is that there are four very long direct repeat sequences (fig.1). These sequences may be involved in enhancing the expression of the rice glutelin gene as shown in the soybean β-conglycinin gene [11], although their functional significance remains unknown. It has been shown that the 5′-flanking regions of leguminous globulin genes are conserved as well as their coding regions. In particular, a long common sequence, designated the legumin box, can be found around 100 bp upstream from the transcription start site [9]. When the legumin box sequence was searched for the 5′-flanking sequence of the rice glutelin gene, a homologous sequence (13/19 bp) could be detected around 60 bp upstream from the transcription start site. Considering that the legumin box is also conserved at a similar position in the rice glutelin gene, such a sequence may be specific to the gene grouped into the 11 S globulin type gene and may play an important role in specifying tissue- and developmental specific expression. In the prolamin genes, a common sequence,

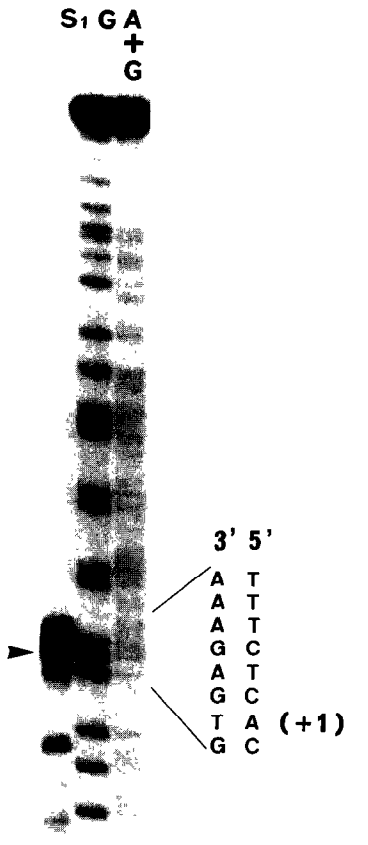



Fig.2. Determination of the transcription start site by $S_1$ mapping. The coding strand of the 343 bp *Hin*fI-*Taq*I fragment, which was labelled at the 5′-end with $^{32}P$, was hybridized with endosperm poly(A) mRNA. The DNA-RNA hybrid was digested with $S_1$ nuclease. The $S_1$-resistant segment (S1) was fractionated using 6% polyacrylamide gel in parallel with the sequence ladders of the coding strand (G and G+A).

named the −300 element, has been found about 300 bp upstream from the start codon [12]. This sequence may be comparable to the legumin box. Therefore, each group of storage protein genes has a unique element in the 5′-flanking region, which may be involved in the control of specific gene expression.

The 3′-end of the glutelin gene was identified by comparison with the type II glutelin cDNA sequence. It has been shown that heterogeneity of the poly(A) addition site can be detected in the type II glutelin gene [1]. Polyadenylation occurs at positions 1947 and 2043. Polyadenylation signals are characteristic of plant genomic genes.

## ACKNOWLEDGEMENT

This work was supported in part by a project grant from the Ministry of Agriculture, Forestry and Fisheries (Japan).

## REFERENCES

[1] Takaiwa, F., Kikuchi, S. and Oono, K. (1987) Mol. Gen. Genet., in press.
[2] Takaiwa, F., Kikuchi, S. and Oono, K. (1986) FEBS Lett. 206, 33-35.
[3] Benton, W.D. and Davis, R.W. (1977) Science 196, 180-182.
[4] Messing, J. (1983) Methods Enzymol. 101, 20-78.
[5] Maxam, A. and Gilbert, W. (1980) Methods Enzymol. 65, 449-559.
[6] Takaiwa, F. and Sugiura, M. (1982) Eur. J. Biochem. 124, 13-19.
[7] Brown, J.W.S. (1986) Nucleic Acids Res. 14, 9549-9559.
[8] Lycett, G.W., Croy, R.R., Shirsat, A.H. and Boulter, D. (1984) Nucleic Acids Res. 12, 4493-4506.
[9] Bäumlein, H., Wobus, U., Pustell, J. and Kafatos, F.C. (1986) Nucleic Acids Res. 14, 2707-2720.
[10] Dunsmuir, P. (1985) Nucleic Acids Res. 13, 2503-2518.
[11] Chen, Z.-L., Schuler, M.A. and Beachy, R.N. (1986) Proc. Natl. Acad. Sci. USA 83, 8560-8564.
[12] Forde, B.G., Heyworth, A., Pywell, J. and Kreis, M. (1985) Nucleic Acids Res. 13, 7327-7339.