

Glycinin A₅A₄B₃ mRNA: cDNA cloning and nucleotide sequencing of a splitting storage protein subunit of soybean

Takayuki MOMMA¹, Takaharu NEGORO¹, Hisashi HIRANO², Akiyo MATSUMOTO¹, Kyoko UDAKA¹ and Chikafusa FUKAZAWA¹

¹ Genetic Engineering Laboratory, National Food Research Institute, Ministry of Agriculture, Forestry and Fisheries, Ibaraki

² National Institute of Agrobiological Resources, Ministry of Agriculture, Forestry and Fisheries, Ibaraki

(Received January 14/March 28, 1985) — EJB 85 0025

The complete nucleotide sequence of a cloned cDNA, designated pGA₅A₄B₃822, corresponding to glycinin A₅A₄B₃ mRNA was determined. Analysis of the cDNA insert revealed that it contained 1899 nucleotides of mRNA sequence with a 5'-terminal non-translated region of 31 nucleotides, a signal peptide region corresponding to 23 amino acids, an acidic subunit region (A₅) corresponding to 97 amino acids, an acidic subunit region (A₄) corresponding to 257 amino acids followed by a basic subunit region (B₃) corresponding to 185 amino acids, and a 3'-terminal non-translated region of 182 nucleotides. These results show that the glycinin A₄ subunit, which is not found to be linked to a basic subunit via a disulfide bond, is synthesized as a full-sized precursor, i.e. the A₅A₄B₃ subunit complex, from a single mRNA, followed by post-translational processing to generate an intermediary subunit complex (A₅-B₃), covalently linked by a disulfide bond, and the mature A₄ subunit, which may associate with the above subunit complex by non-covalent interactions. From the results obtained by the Chou-Fasman rules we speculated that the two post-translational cleavage sites of this subunit precursor might be processed by the same proteolytic enzyme.

Glycinin, one of the principal storage proteins of soybean (*Glycine max* L., Merr.), is composed of six subunit pairs, each pair consisting of an acidic and a basic polypeptide that are linked via a disulfide bond, in one molecule [1–3]. The subunit pairings of glycinin molecules are themselves a heterogeneous complex of A_{1a}B₂, A_{1b}B_{1b}, A₂B_{1a}, A₃B₄ and A₅B₃, which are covalently linked by disulfide bonds unlike the A₄ subunit [4, 5].

Recently Marco et al. have determined the partial genomic structure for a glycinin A₂B_{1a} subunit precursor [6]. We have also sequenced the double-stranded (ds) cDNA encoding a glycinin precursor for the A₃ acidic and the B₄ basic subunits [7]. These results conclusively showed that the specific acidic and basic components of every glycinin subunit pair are synthesized from the same mRNA as a single precursor, which may undergo a posttranslational processing event in order to form the individual linked subunit as speculated by Tumer et al. [8].

A question is now raised regarding the origin of a glycinin acidic subunit, A₄, which is not covalently linked to any basic subunits [5].

In this communication, we report that the A₄ subunit originates from a precursor of the A₅A₄B₃ subunit complex, synthesized from a single mRNA by specific proteolytic processing. The deduced complete amino acid sequences of the A₅, A₄ and B₃ subunits have also been determined.

EXPERIMENTAL PROCEDURE

Harvesting and storage of soybean cotyledons

Soybean cultivar Bonminori (*Glycine max* L., Merr.) was used in this experiment as described previously [7].

Correspondence to C. Fukazawa, Genetic Engineering Laboratory, National Food Research Institute, Ministry of Agriculture, Forestry and Fisheries, 2-1-2 Kannondai, Yatabe-Machi, Tsukuba-gun, Ibaraki-ken, Japan 305

Abbreviations. bp, base pairs; ds, double stranded.

Preparation of RNA

Total RNA was extracted from pulverized frozen tissue with phenol/sodium dodecyl sulphate [7, 9]. Poly(A)-containing RNA was prepared by oligo(dT)-cellulose affinity chromatography [10].

Construction of a soybean cotyledonary ds-cDNA recombinant plasmid library

Double-stranded cDNA was synthesized from total poly(A)-containing RNA, sized, tailed, inserted and then transfected into *Escherichia coli* RR1 or *E. coli* GM 2150 as described previously [7, 11, 12]. The recombinant plasmid library was established by a successive selection using tetracycline resistance and colony hybridization to ³²P-labeled cDNA prepared from the total poly(A)-containing RNA. All the recombinant work was performed under containment conditions in accordance with the Genetic Manipulation Advisory Group (Japan) guidelines.

Identification of transformants containing the A₅ subunit sequence

About 1500 individual transformants of the soybean recombinant plasmid library were screened by colony-DNA filter hybridization to a ³²P-labeled 51-base-pair (bp) *Sph*I-*Mlu*I restriction endonuclease DNA fragment according to the procedure described by Grunstein and Hogness [13] (see Fig. 1).

Plasmid DNA preparation and restriction analysis

Plasmid DNA was isolated and purified as described previously [7]. Restriction enzymes were purchased from Nippon Gene Inc. and New England Biolabs.

Nucleotide sequencing

Nucleotide sequences were determined by the methods of Maxam and Gilbert [14] and examined by computer analysis (Software Development Corporation, Japan).

RESULTS

Isolation and characterization of the cloned cDNA for glycinin A₅A₄B₃ subunit precursor

From the soybean cotyledonary cDNA library six colonies that showed an intensive hybridization signal with the ³²P-labeled *Sph*I-*Mlu*I DNA fragment, as demonstrated in Fig. 1, were selected in order to determine the partial nucleotide sequences. As a result of these structural analyses a recombinant plasmid containing a ds-cDNA insert of about 2000 bp was identified as the plasmid containing both the A₅ and the B₃ subunit sequences and was designated as pGA₅A₄B₃822.

The complete nucleotide sequence of this cloned glycinin A₅A₄B₃ subunit ds-cDNA was determined. The restriction endonuclease fragments and the sequencing strategy used in the sequence determination of the ds-cDNA are indicated in Fig. 2. As shown in Fig. 3, the overall length of the sequence in this cloned cDNA is 1899 nucleotides, corresponding to glycinin A₅A₄B₃ mRNA.

At the 5' terminus a non-translated region of 31 nucleotides, preceding the AUG translation start codon, was found in the cloned cDNA. The isoleucyl residue that marks the NH₂ terminus of mature A₅ subunit [4] is specified by the twenty-third codon following the AUG initiation codon. The preceding 23 codons, which are identical to the corresponding codons of A₃B₄ cDNA except for the lack of a phenylalanyl residue (sixth codon following the initiation codon), appear to encode a signal peptide, characteristic for secreted proteins [15], which is probably removed during cotranslational processing [16]. Analysis of the coding region of glycinin A₅A₄B₃ subunit mRNA indicated that these A₅, A₄ and B₃ subunits were synthesized as a single precursor polypeptide, which underwent posttranslational processing to form an A₅-B₃ intermediary subunit complex, covalently linked by a disulfide bond but with non-covalent association with the A₄ subunit. The mature A₅ subunit is encoded by 291 nucleotides (97 amino acids) and the mature A₄ and B₃ subunits are encoded by 771 nucleotides (257 amino acids) and 555 nucleotides (185 amino acids) respectively in the A₅A₄B₃ mRNA, as the positions of the NH₂-terminal and the COOH-terminal residues (unpublished results) of those mature subunits were identified by a comparison to the partial amino acid sequences of those subunits [4, 17]. The relative molecular masses, calculated from the inferred amino acid sequences of A₅, A₄ and B₃ subunits, were 10556, 29967, and 20732 respectively. Following the UAA translation stop codon is a 3'-terminal non-translated region of 182 nucleotides, adjacent to a poly(A) segment. The sequence AAUAAA, usually found near the 3' end of eukaryotic mRNAs [18], is located in the glycinin A₅A₄B₃ mRNA, ending 25 nucleotides upstream from the poly(A) segment. This unique sequence is also found in an unexpected position 103 nucleotides from the end of the segment. The 3' non-translated region of this mRNA contained the sequence AAUAAAUAUAAA, which could be observed in those of pea legumin [19] and soybean legheamoglobin [19], but 119 nucleotides from the poly(A) segment.

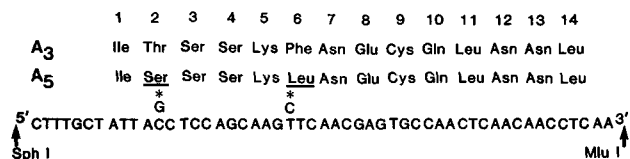


Fig. 1. Nucleotide sequence and the inferred amino acid sequence of *Sph*I-*Mlu*I DNA fragment derived from the cDNA insert of plasmid pGA₃B₄1425 [7] and used to probe the soybean cotyledonary cDNA library. The possible single-base mutation in the DNA, necessary to accommodate the amino acid changes in the A₅ subunit sequence, is also indicated by *

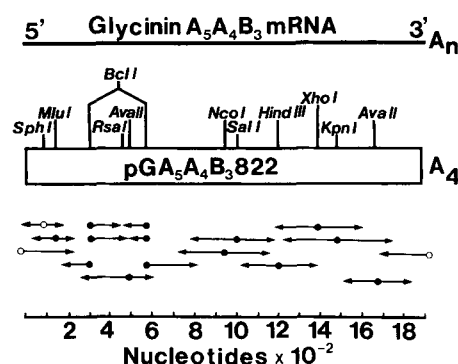


Fig. 2. Restriction endonuclease map and nucleotide sequencing strategy for glycinin A₅A₄B₃ subunit ds-cDNA insert of pGA₅A₄B₃822. The top horizontal line represents the length of A₅A₄B₃ mRNA corresponding to the relative positions of the cloned ds-cDNA. The cloned ds-cDNA is shown with only those restriction endonuclease sites employed in the sequence determination. Each of the ends of the ds-cDNA is bound by segments containing about 18 G·C residues and also contains a coding-strand oligo(dA) segment of four nucleotides at the 3' end. Arrows indicate the direction and length of sequencing of restriction endonuclease-digested DNA fragments. (●) The sites labeled at flush or recessed 5' ends; (○) the sites labeled at protruding 3' ends

Comparison of nucleotide sequences of the A₅A₄B₃ and the A₃B₄ cDNAs

An extensive homology occurs in the A₃ family mRNAs of glycinin (see Fig. 3 and [7]). To achieve an alignment with maximum homology, hypothetical gaps were introduced at thirteen regions of the A₃B₄ mRNA sequence, and were introduced at fourteen regions of the A₅A₄B₃ mRNA sequence (data not shown). With this nucleotide alignment, it was found that overall, 82% of the nucleotide positions in the A₃ family mRNAs were identical. Both nucleotide sequences in the signal peptide region show an identity in 96% of the nucleotide positions, while those in the determined sequence of the entire 5' non-translated region and in 3' non-translated region of those mRNAs show the same identity in 80% of the positions. A comparison of the nucleotide positions in those mRNAs also revealed that 91% of the positions in the coding sequences for mature A₅ subunit and the A₅ corresponding region of mature A₃ subunit, and 78% of those for mature A₄ subunit and the corresponding region of that protein, were identical to each other. On the other hand, both nucleotide sequences in the region of the basic subunits, B₃ and B₄, show an identity in 82% of the nucleotide positions. The relatively long insertion, which corresponded to 30 amino acids (90 nucleotides),

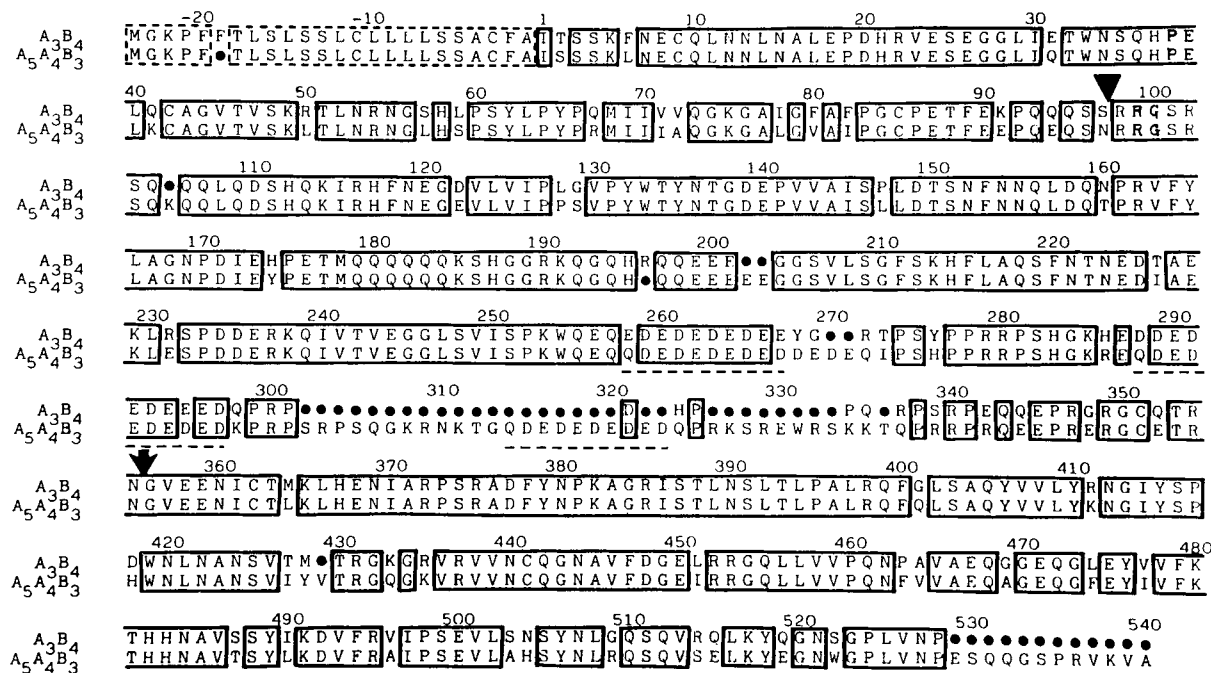


Fig. 4. Comparison of the amino acid sequence of glycinin $A_5A_4B_3$ subunit with that of glycinin A_3B_4 subunit. The amino acid sequence of the A_3B_4 subunit precursor is cited from [7]. The single-letter notation for amino acids is in accordance with the recommended convention of Dayhoff [26]. The dashed boxes enclose amino acids that are identical in the signal peptide segments of the two glycinin subunits, whereas the solid boxes enclose the identical amino acids of those precursor polypeptides. (●) (at several positions) The hypothetical deletion spaces that were placed in the each sequence to maximize the homology between the two glycinin subunits. (▼) The cleavage site, which produces the mature A_5 and A_4 subunits during a posttranslational processing (while the A_3B_4 subunit precursor does not undergo such processing step in the corresponding region). (⇓) The cleavage sites between acidic and basic subunits. (---) The repeated units

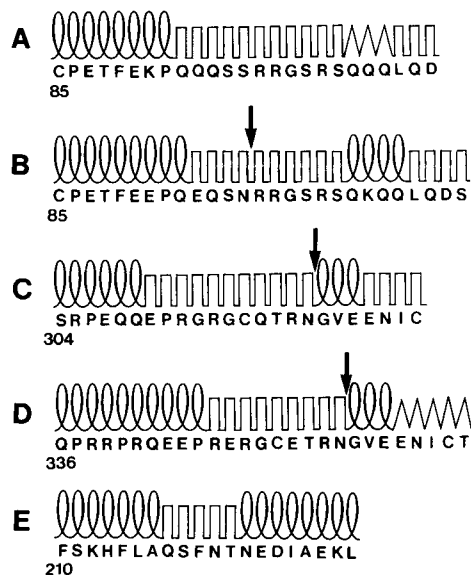


Fig. 5. Schematic representation of the predicted secondary structure around the posttranslational cleavage sites in the $A_5A_4B_3$ subunit precursor and around the corresponding regions of the A_3B_4 subunit precursor. Loops, α helix, β turn, β -pleated sheet. Each symbol represents one amino acid residue. The numbers below the drawings represent the amino acid positions beginning the specific domain structure in the $A_5A_4B_3$ and the A_3B_4 subunit precursors. The single-letter notation for amino acids is employed [26]. The arrows indicate the cleavage site of each amino acid sequence. (A) The predicted secondary structure for the region spanning residues 85–109 in the A_3B_4 subunit precursor; (B) residues 85–111 in the $A_5A_4B_3$ subunit precursor; (C) residues 304–327 in the A_3B_4 subunit precursor; (D) residues 336–362 in the $A_5A_4B_3$ subunit precursor; (E) residues 210–229 in the $A_5A_4B_3$ subunit precursor

proteins of the pea [25], the result was not in agreement with this rule (data not shown). This may suggest that the processing enzymes are different in both storage proteins. On the other hand, the segment which was predicted to have the α -helix (more than two α -helical turns predicted) – β -structure (asparagine contained) – α -helix (more than two helical turns predicted) was not cleaved in the posttranslational processing event (Fig. 5E).

DISCUSSION

Since it has been established that a specific pair of acidic and basic subunits of glycinin is synthesized from a single mRNA [6, 7], our interest was focused on the A_5B_3 subunit complex, which was rather shorter than the other glycinin A-B subunit complexes, covalently linked by disulfide bonds, and also on the A_4 subunit, the polypeptide not covalently linked to another subunit [5]. Although the apparent molecular masses of $A_5 + B_3$ and A_4 subunits from sodium dodecyl sulphate electrophoresis are about 32000 and 35000 Da, respectively, Northern blotting, using the DNA fragment probe for the A_3 subunit family, indicated that the reactive mRNA species was only around 2000 nucleotides long [7]. Moreover, the internal amino acid sequence, which began with residue Arg-98 of the A_3 subunit was highly homologous to the NH_2 -terminal sequence of the A_4 subunit [17, 28], whereas the A_5 NH_2 -terminal sequence was also highly homologous to the A_3 NH_2 -terminal sequence [4]. From these results it should be supposed that both the A_4 subunit and the A_5B_3 subunit complex are derived from a full-sized precursor having $M_r \approx 67000$. Recently Staswick et al. have proposed the same hypothesis for the origin of the A_4 subunit [29]. To confirm

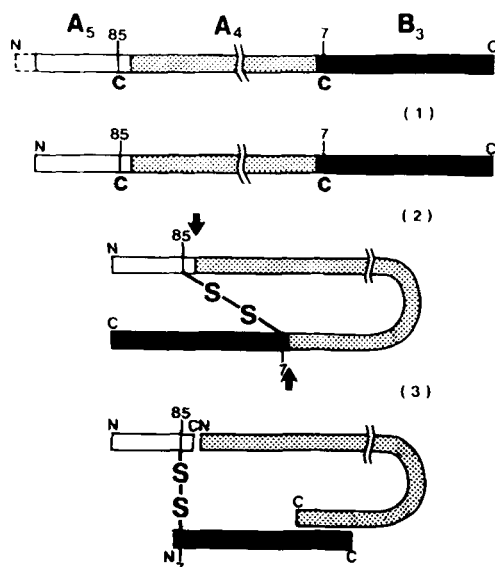


Fig. 6. Schematic drawing of the proteolytic processing steps in glycinin $A_5A_4B_3$ subunit precursor. (—) A signal peptide region in that precursor; (□) The A_5 subunit region; (▨) the A_4 subunit region of the precursor; (■) the B_3 subunit region. N and C, the NH_2 terminus and the $COOH$ terminus of each protein respectively. C, the positions of the cysteine residues, which form a disulfide bond between the A_5 and the B_3 subunits. (↓) The cleavage sites during the post-translational processing

these speculations the partial nucleotide sequencing of the cDNA clones selected from the soybean cDNA library with a unique probe (see Fig. 1) was performed. As the amino acid sequence, deduced from the partial nucleotide sequence of pGA $_5A_4B_3$ 822, was completely identical to those of the NH_2 -terminal regions of both the A_5 and the B_3 subunits [4], the cDNA insert of this clone was completely sequenced and the inferred amino acid sequence was also demonstrated (Fig. 3). Comparison of the deduced complete amino acid sequence with the NH_2 -terminal regions of the mature A_5 , A_4 and B_3 subunits [4] revealed that these three subunits are synthesized as a single precursor protein. The determination of the $COOH$ -terminal residues of the mature A_5 , A_4 and B_3 subunits, according to the carboxypeptidase A analyses of CNBr peptides of their subfragments (data not shown), revealed that there was no evidence for a linker sequence joining the acidic and the basic subunits [6, 7, 19].

The striking similarities between the nucleotide sequences of the $A_5A_4B_3$ and the A_3B_4 subunit mRNAs [7] as well as the amino acid sequence between both proteins (Fig. 4) have major implications for evolutionary rearrangement of glycinin subunit genes, as well as for the structures essential to the maturation of the subunit proteins. Nevertheless the very high degree of sequence conservation occurred in both precursor proteins of glycinin; the $A_5A_4B_3$ subunit precursor contained an additional posttranslational cleavage site between the amino acid positions 97 (asparagine) and 98 (arginine), as well as the cleavage site between the acidic and the basic subunits, which might exist commonly in every glycinin subunit precursor. This is the most outstanding feature of this precursor.

It is of importance to know why only this precursor can be cleaved in the acidic polypeptide region among the glycinin subunit precursors during the posttranslational processing. In order to answer this question we also applied the rules for

empirical prediction of protein secondary structure [22, 23] to the region spanning residues 85–111 of this polypeptide. The result indicated that the predicted structure around the cleavage site of the acidic polypeptide region was quite similar to those between the acidic and the basic subunits of the glycinin and the legumin (the pea storage protein, which is analogous to glycinin) precursors.

By comparing the structure of this acidic polypeptide region with that of the corresponding region of the A_3B_4 subunit precursor, it might be considered that the conformational change around the acidic cleavage site was based on the insertion of a lysine at the amino acid position 105, accompanied by the replacement of a serine with an asparagine at position 97 (Figs 4 and 5).

The conservation of the relatively striking hydrophilicity, which could be observed in the $COOH$ -terminal regions of the acidic subunits of the glycinin and the legumin, may be at least as important as the conservation of charge in order to retain the non-covalent interactions with the basic counterparts besides the covalent linkage between both subunits. Therefore, the glycinin A_4 subunit needs to have a higher degree of potential for interaction with the B_3 subunit. Thus we supposed that both the third 10-amino-acid repeated unit, located in the region spanning residues 313–323, and the 12-amino-acid segment, located in the $COOH$ terminus of the $A_5A_4B_3$ subunit precursor, might contribute to strengthening the potential of the non-covalent interactions of the A_4 subunit with the A_5 - B_3 subunit complex.

As the position of a disulfide bond in the A_5 - B_3 subunit complex was determined recently [26], the proteolytic processing of the $A_5A_4B_3$ subunit precursor can be represented diagrammatically as shown in Fig. 6. In step 1, preceding the isoleucine residue is a 23-amino-acid segment, beginning with methionine, which may function as a signal peptide and is probably removed during cotranslational processing. In step 2 the $A_5A_4B_3$ subunit precursor may undergo posttranslational processing to form a disulfide bond between the A_5 subunit region and the B_3 subunit region, followed by the proteolytic digestion at the two cleavage sites shown by the short arrows. The processing enzyme may recognize the specific domain structure and cleave at the end of the asparagine residue. In step 3 a consecutive conformational change occurs in the three subunits in order to retain the stable structure of the glycinin subunit complex.

We are grateful to Dr Kyuya Harada for his help in breeding the soybean plants. We also appreciate the excellent technical assistance of Yukie Isoyama, and the manuscript preparation and editorial assistance of Noriko Kobayashi. T. M. is a research fellow from Kirin Brewery Co., Plant Research Center, Kitsuregawa, Tochigi and recipient of a predoctoral trainee fellowship from National Food Research Institute. T.N. is a research fellow from Sumitomo Chemical Co., Takarazuka Research Center, Takarazuka, Hyogo, and recipient of a predoctoral trainee fellowship from National Food Research Institute.

REFERENCES

1. Catsimopoulos, N., Kenney, J. A., Meyer, E. W. & Szuhaj, B. F. (1971) *J. Sci. Food Agric.* 22, 448–453.
2. Kitamura, K., Takagi, T. & Shibasaki, K. (1977) *Agric. Biol. Chem.* 41, 833–840.
3. Moreira, M. A., Hermodson, M. A., Larkins, B. A. & Nielsen, N. C. (1981) *Arch. Biochem. Biophys.* 210, 633–642.
4. Tumer, N. E., Richer, J. D. & Nielsen, N. C. (1982) *J. Biol. Chem.* 257, 4016–4018.

5. Staswick, P. E., Hermodson, M. A. & Nielsen, N. C. (1981) *J. Biol. Chem.* 256, 8752–8755.
6. Marco, Y. A., Thanh, V. H., Tumer, N. E., Scallan, B. J. & Nielsen, N. C. (1984) *J. Biol. Chem.* 259, 13436–13441.
7. Fukazawa, C., Momma, T., Hirano, H., Harada, K. & Udaka, K. (1985) *J. Biol. Chem.*, in the press.
8. Tumer, N. E., Thanh, V. H. & Nielsen, N. C. (1981) *J. Biol. Chem.* 256, 8756–8760.
9. Brawerman, G., Mandecki, J. & Lee, S. Y. (1972) *Biochemistry* 11, 637–641.
10. Aviv, H. & Leder, P. (1972) *Proc. Natl Acad. Sci. USA* 69, 1408–1412.
11. Land, H., Grez, M., Hauser, H., Lindenmaier, W. & Schutz, G. (1981) *Nucleic Acids Res.* 9, 2251–2266.
12. McLean, J. W., Fukazawa, C. & Taylor J. M. (1983) *J. Biol. Chem.* 258, 8993–9000.
13. Grunstein, M. & Hogness, D. S. (1975) *Proc. Natl Acad. Sci. USA* 72, 3961–3965.
14. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* 65, 499–560.
15. Kreil, G. (1981) *Annu. Rev. Biochem.* 50, 317–348.
16. Lin-Lee, Y.-C., Bradley, W. A. & Chan, L. (1981) *Biochem. Biophys. Res. Commun.* 99, 654–661.
17. Hirano, H., Fukazawa, C. & Harada, K. (1984) *J. Biol. Chem.* 259, 14371–14377.
18. Benoist, C., O'Hare, K., Breathnach, R. & Chambon, P. (1980) *Nucleic Acids Res.* 8, 127–142.
19. Lycett, G. W., Delauney, A. J., Zhao, W., Gatehouse, J. A., Croy, R. R. D. & Boulter, D. (1984) *Plant Mol. Biol.* 3, 91–96.
20. Hyldig-Nielsen, J. J., Jensen, E. O., Paludan, K., Wiborg, O., Garret, R., Jorgensen, P. & Marcker, K. A. (1982) *Nucleic Acids Res.* 10, 689–701.
21. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* 157, 105–132.
22. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* 13, 211–222.
23. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* 13, 223–245.
24. Schulz, G. E. & Schirmer, R. H. (1979) *Principles of protein structure* pp. 69–70, Springer, New York.
25. Gatehouse, J. A., Lycett, G. W., Delauney, A. J., Croy, R. R. D. & Boulter, D. (1983) *Biochem. J.* 212, 427–432.
26. Dayhoff, M. O. (ed.) (1972) *Atlas of protein sequence and structure*, vol. 5, pp. D2–D6, National Biochemical Research Foundation, Washington, DC.
27. Staswick, P. E., Hermodson, M. A. & Nielsen, N. C. (1984) *J. Biol. Chem.* 259, 13431–13435.
28. Hirano, H., Fukazawa, C. & Harada, K. (1985) *FEBS Lett.* 181, 124–128.
29. Staswick, P. E., Brouè, P. & Nielsen, N. C. (1983) *Plant Physiol.* 72, 1114–1118.
30. Lycett, G. W., Croy, R. R. D., Shirsat, A. H. & Boulter, D. (1984) *Nucleic Acids Res.* 12, 4493–4506.