

Cloning of cDNA encoding the sweet-tasting plant protein thaumatin and its expression in *Escherichia coli*

(Recombinant DNA; preprothaumatin; molecular cloning; signal sequence; processing of carboxy terminus)

Luppo Edens*, Lammert Heslinga**, Rob Klok**, Adrianus M. Ledebøer*,
Jan Maat*, Marjolein Y. Toonen*, Chris Visser* and C. Theo Verrips*

* Section of Microbiology, and ** Section of Organic Chemistry, Unilever Research Laboratory, P.O. Box 114, 3130 AC Vlaardingen (The Netherlands)

(Received October 1st, 1981)

(Accepted December 12th, 1981)

SUMMARY

The structural gene of the sweet-tasting plant protein (prepro)thaumatin was cloned and expressed in *Escherichia coli*. Expression was effected under control of *lac* and *trp* promoter/operator systems and through the use of bacterial ribosome-binding sites. The naturally occurring thaumatin II represents a processed form. The primary translation product, preprothaumatin, of the cloned mRNA-derived cDNA contains extensions at both the amino terminus and the carboxy terminus. The amino terminal extension of 22 amino acids is hydrophobic and very much resembles an excretion-related signal sequence. The six amino acids-long carboxy terminal extension is very acidic in character, in contrast to the overall highly basic thaumatin molecule. The possible role of such an acidic tail with respect to compartmentalization is discussed.

INTRODUCTION

Of late there has been a growing interest in sweet substances that can replace sucrose. The rationale to use such compounds is the wish to reduce the cariogenicity or diabetogenicity of a

product or to lower its calorific value. In the last few years attention has therefore been focussed on some extremely sweet proteins present in the fruits of several tropical plants. Among these proteins, thaumatin seems the most promising.

Thaumatin is isolated from the arils of the fruits of *Thaumatococcus daniellii* Benth, a West African shrub. At least five different forms of thaumatin can be isolated by ion-exchange chromatography. All these forms are nearly 100000 times sweeter than sucrose on a molar base and they are called thaumatin I, II, III (Van der Wel and Loeve, 1972), b and c (Higginbotham and Hough, 1977). Their M_r -values are about 22000

Abbreviations: aa, amino acid(s); bp, base pairs; PAS, *p*-aminosalicylic acid; RuBPCase, ribulose biphosphate carboxylase; SDS, sodium dodecyl sulfate; STDS, 0.9% NaCl, 1% Triton X-100, 0.5% sodium deoxycholate and 0.5% SDS.

and all are basic proteins having isoelectric points of almost pH 12. Thaumatin I consists of a single polypeptide chain of 207 amino acid residues, and its primary amino acid sequence is known (Iyengar et al., 1979). The two predominant thaumatin I and II vary in quantity depending on the batch of fruits tested, probably reflecting genetic variation among strains as fruits are collected at random from non-cultivated plants.

Because of their extreme sweetness, the thaumatinins are suitable probes in taste mechanism studies (Glaser et al., 1978). In this respect it is noteworthy that, besides man, only Old World monkeys give a positive taste response. The structural requirements determining the sweet character of the thaumatinins are investigated both by means of chemical modification and by analysis of their tertiary structure (Van der Wel and Bel, 1976). An important conclusion from the latter studies is that the positive charge of some of the lysine residues is necessary for eliciting the sweet taste sensation.

As yet nothing is known about the regulation of synthesis and processing of thaumatin in the fruit. To obtain information on these phenomena and on the genetic organisation of the thaumatinins, the complete mRNA-derived cDNA of thaumatin was cloned. Furthermore, the availability of such a clone and knowledge of its primary nucleotide sequence will allow the introduction of specific variations in the protein through in vitro site-directed mutagenesis. Together with the chemical modification studies such modified thaumatinins might have important implications for our understanding of the perception of the sweet taste.

So far only a few structural plant genes have been cloned. Determination of their primary nucleotide sequences revealed that these clones contain only part of the coding information of the genes. Recently it was reported that the plant-related gene for the large subunit of ribulose biphosphate carboxylase of maize and wheat chloroplasts has been brought to expression in *E. coli* (Gatenby et al., 1981).

In this article the molecular cloning and the nucleotide sequence of the genetic information for the sweet plant protein thaumatin are described. In addition, synthesis of the thaumatin II precursor in *E. coli* is demonstrated.

MATERIALS AND METHODS

(a) Bacteria and nucleic acids

E. coli K-12, strain HB101 and *E. coli* K-12, strain 294 were obtained from H.W. Boyer and K. Backman. Plasmid DNA was isolated by the procedure described by Birnboim and Doly (1979).

(b) Enzymes

All restriction enzymes used were purchased either from Boehringer, New England Biolabs or Bethesda Research Labs, except restriction enzyme *RruI* which was a gift from A. de Waard (University of Leiden). Terminal transferase and *E. coli* DNA polymerase I-large fragment were obtained from Bethesda Research Labs, T4-polynucleotide kinase and *E. coli* DNA polymerase I from Boehringer and nuclease S1 and lysozyme from Sigma. *E. coli* ATP:RNA adenylyl transferase was purified as described by Sippel (1973). T4-DNA ligase was isolated according to the procedure described by Murray et al. (1979) except for the final purification step in which use was made of hydroxyapatite chromatography.

(c) Purification and radioactive labeling of RNA

Arils were separated from *Thaumatococcus* fruits and stored under liquid nitrogen. Frozen arils were powdered mechanically and extracted with buffer (10 mM Tris · HCl, pH 8.0, 5 mM MgCl₂, 1% NaCl, 6% PAS, and a mixture of phenol, chloroform and isoamylalcohol (25:24:1)). The extracted nucleic acids were recovered by ethanol precipitation. From the dissolved nucleic acids RNA was selectively precipitated by the addition of 0.5 vol. of 8 M LiCl and further purified by Sephadex G-75 gel filtration. Polyadenylated RNA was isolated by two passages over oligo(dT) cellulose columns. 10 g of arils yielded approx. 3 mg total RNA and 15 µg polyadenylated RNA. RNA was radioactively labeled with ATP:RNA adenylyl transferase using conditions described by Sippel (1973). To prevent synthesis of long poly(A) tails, the amounts of RNA and [α -³²P]ATP added were chosen such that an average of less than one ATP residue per RNA molecule was incorporated. Con-

trol experiments showed that under the conditions used less than 0.1% of radioactively labeled ribosomal RNA was retained by oligo(dT) cellulose chromatography. For preparative isolation of thaumatin-encoding mRNA, 20 µg polyadenylated RNA was mixed with 2 µg of a similar batch of α -³²P-labelled RNA. The mixture was then subjected to electrophoresis through a 2.5% polyacrylamide gel in the presence of 7 M urea. Electrophoresis was performed in E-buffer (40 mM Tris · acetate pH 7.8, 20 mM sodium acetate, 2 mM EDTA) supplied with 0.1% SDS. After autoradiography radioactive bands were eluted from the polyacrylamide gel and analysed for their thaumatin-coding capacity in a wheat germ-derived protein synthesizing system.

(d) Protein synthesis and analysis of translation products

Cell-free protein synthesis in a wheat germ-derived system was accomplished as described by Davies and Kaesberg (1973). The translation products were analysed by electrophoresis through 13.5% polyacrylamide gels according to Wyckoff et al. (1977). Denaturation of the protein samples was achieved by heating the samples for 5 min at 100°C in sample buffer (20 mM Tris · HCl, pH 6.8, 2% SDS, 5% β -mercaptoethanol, 6 M urea and bromophenol blue).

(e) Double-stranded cDNA synthesis

Messenger RNA eluted from the polyacrylamide gel was converted into single-stranded cDNA according to Buell et al. (1978). The addition of 4 mM sodium pyrophosphate to the reaction mixture was found to considerably stimulate the formation of long transcripts. The second DNA strand was synthesized as described by Wickens et al. (1978) except that we used DNA polymerase large fragment instead of DNA polymerase I. After phenol extraction of the reaction mixture the aqueous layer was passed over a Sephadex G-50 column equilibrated with nuclease S1 digestion buffer (400 mM NaCl, 200 mM Na · acetate, 2.5 mM ZnCl₂, pH 4.5); subsequently, the DNA in the excluded fractions was treated with nuclease S1 at a concentration of 200 units/ml at 15°C for

30 min. The nuclease S1-treated material was then subjected to centrifugation through a 5 – 30% glycerol gradient in STE (100 mM NaCl, 10 mM Tris · HCl, pH 7.6, 1 mM EDTA). Double-stranded DNA fragments longer than 750 nucleotides were selected and elongated with dC residues using terminal transferase. These fragments were then annealed into *Pst*I-cleaved pBR322 which had been tailed with about 10 dG residues.

(f) Transformation and identification of clones

E. coli HB101 was transformed with the recombinant DNA molecules. From the transformants 10 colonies were chosen arbitrarily and from these colonies about 10 µg of plasmid DNA was isolated (Birnboim and Doly, 1979). These plasmid DNA preparations were cleaved with *Eco*RI, denatured and fixed to DBM paper (1 cm²). After preincubation each of the DNA filters was hybridized for 6 h to 5 µg of polyadenylated *Thaumatococcus* RNA (Williams et al., 1979). The filters were washed extensively and the retained RNA was eluted, ethanol-precipitated and translated in a wheat germ protein-synthesizing system.

(g) Nucleotide sequence analysis

DNA sequences were determined both by the chemical procedure of Maxam and Gilbert (1980) and the enzymatic dideoxy/nick translation procedure of Maat and Smith (1978). RNA sequence analysis was performed with the primed-synthesis approach using AMV reverse transcriptase in the presence of dideoxynucleoside triphosphates. The protocol used was essentially as described by Hamlyn et al. (1978). The double-stranded *Hae*III-*Taq*I-fragment extending from positions 36-105 was converted into single-stranded DNA by exonuclease III. This single-stranded DNA (1 pmol) was mixed with 2 µg polyadenylated *Thaumatococcus* RNA to prime cDNA synthesis.

(h) Construction of expression plasmids

Plasmid pUR201, a pBR322 derivative containing two *lacUV* 5 promoters/operators with one *Eco*RI-site, was constructed essentially as described for pGH6 by Goeddel et al. (1979). Plas-

mid pUR301, also a pBR322 derivative, contains the tryptophan promoter/operator region but lacks the attenuator region. This plasmid was constructed as follows. Plasmid *ptrp* ED5-1 (Hallewell and Emtage, 1980) was digested with *TaqI* in the presence of *E. coli* RNA polymerase to protect the *TaqI* site present in the *trp* promoter site. The 230 bp-long fragment containing the intact promoter was then treated with nuclease S1, provided with *EcoRI* linkers (dGGAATTCC) and cloned in the *EcoRI* site of pBR322. This yielded plasmid pUR300. The orientation of the insert was such that the tryptophan promoter gives rise to transcription in the same direction as the tetracycline-resistance gene of pBR322 is transcribed. One of the *EcoRI* sites is located just downstream of the ribosome-binding site of the tryptophan leader peptide (Yanofsky, 1981). The other *EcoRI* site was removed by nuclease S1 treatment and subsequent ligation of partially *EcoRI*-cleaved pUR300. This resulted in plasmid pUR301. Partial cleavage was effected by the presence of 150 $\mu\text{g/ml}$ ethidium bromide. The DNA sequence of the promoter/operator region of pUR301 is shown in Fig. 1. Plasmid pUR105 was constructed by cloning the *PstI* insert of pUR100 in the *EcoRI* site of

pBR322. For this purpose the *PstI* insert was treated with nuclease S1 and then supplied with *EcoRI* linkers (dGGAATTCC). The construction of plasmid pUR521 is delineated in Fig. 2.

Inserts of plasmids pUR521 and pUR110 were ligated into the *RruI*-site at position 600. Cloning of the newly constructed insert in pUR201 and pUR301 resulted in pUR528 and pUR538, respectively. The latter plasmids contained a full-length cDNA copy of thaumatin II mRNA downstream of the ATG-initiation codon, inserted between the *EcoRI* and *HindIII*-site of the expression plasmids.

(i) Antibody precipitation from bacterial extracts

E. coli 294 harbouring either plasmid pUR531, plasmid pUR538 or plasmid pUR301 was grown to an absorbance at 600 nm of 0.1 (5×10^7 cells/ml) in M9-minimal medium supplemented with glucose (0.4%), thiamine (0.2 $\mu\text{g/ml}$), ampicillin (50 $\mu\text{g/ml}$) and all amino acids (50 $\mu\text{g/ml}$) except cysteine. Beta-indolyl acrylic acid was then added to a final concentration of 40 $\mu\text{g/ml}$ to derepress the tryptophan promoter. After 1 h [^{35}S]cysteine (3×10^{11} Bq/mmol) was added to a final concentration of 0.8 μM . Cells were labeled for 5 min at 37°C. Subsequently, a 100-fold excess of unlabelled cysteine was added, the culture (1 ml) was chilled quickly and the cells were collected by centrifugation for 1 min in an eppendorf centrifuge. Bacterial pellets were resuspended in 150 μl 50 mM Tris · HCl pH 8.0, 10 mM EDTA, 0.25 M sucrose and lysozyme at a concentration of 2 mg/ml. After 30 min of incubation on ice the suspensions were subjected to 5 cycles of freezing and thawing. Subsequently, MgCl_2 (final concentration 18 mM), DNase (20 $\mu\text{g/ml}$) and RNase (10 $\mu\text{g/ml}$) were added. Incubation took place for 30 min at room temperature. For direct analysis of the radioactively labelled proteins, 10% of the lysed cells were precipitated directly with acetone and heated for 5 min at 100°C in sample buffer. For immunoprecipitation, the remaining 90% were adjusted to STDS in 10 mM sodium-phosphate buffer pH 7.2 and to each sample 5 μl of rabbit anti-thaumatins serum was added. After incubation for 16 h at 4°C an equal volume (200 μl) of STDS in 10 mM Tris · HCl pH 8.5 was added together

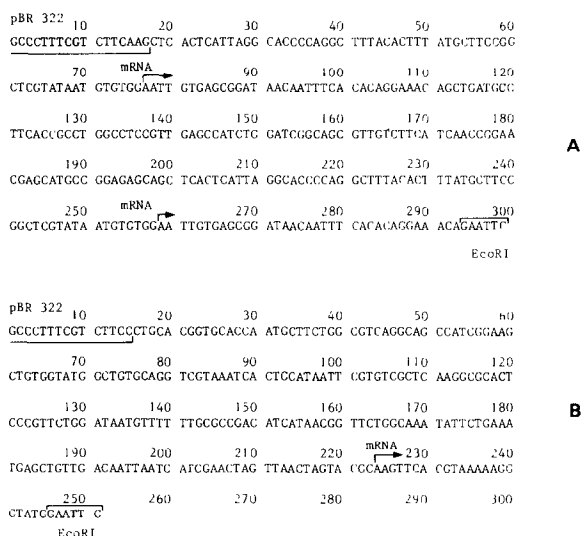


Fig. 1. DNA sequences of the regions covering the double *lacUV5* promoter/operator in pUR201 (A) and the tryptophan promoter/operator in pUR301 (B). The underlined sequence is derived from pBR322 (Sutcliffe, 1979). The starting points of mRNA transcription are indicated (Squires et al., 1976).

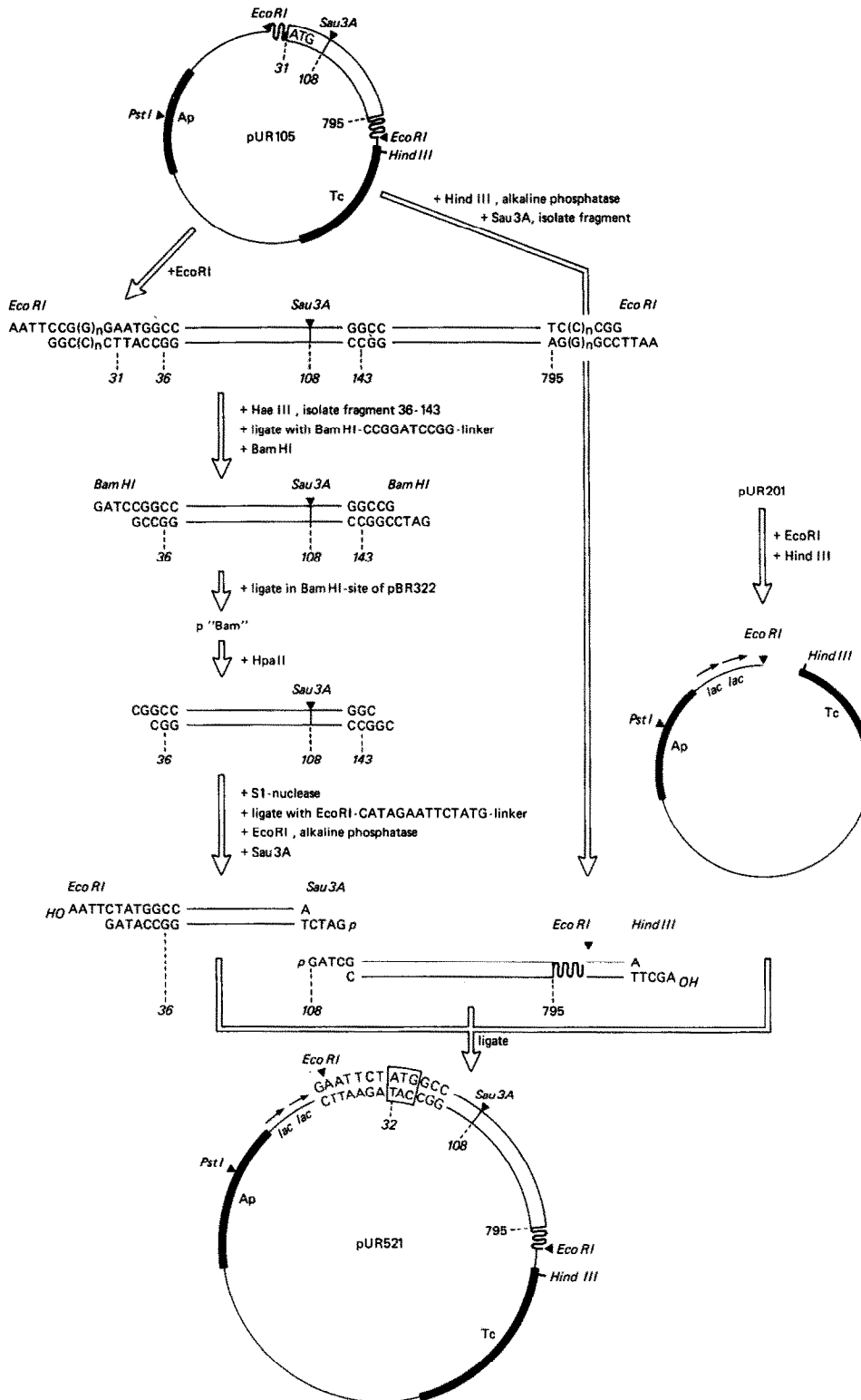


Fig. 2. Construction of plasmid pUR521 directing the synthesis of preprothaumatin in *E. coli* under control of the lactose regulon. Wavy line, dG/dC-stretch; thin line, pBR322-DNA-derived sequences; thick line, ampicillin (Ap) and tetracycline (Tc) resistance-encoding gene sequences; double line, preprothaumatin-encoding sequences. The numbers refer to the DNA sequence given in Fig. 6.

with 50 μ l of a 10% (v/v) suspension of protein A-sepharose (Pharmacia) in STDS. After 30 min at room temperature, the mixture was layered over 300 μ l 10% sucrose in STDS/10 mM Tris · HCl pH 8.2 and centrifuged for 1 min in an eppendorf centrifuge. The liquid phase was discarded and the protein A-sepharose beads were washed three times with STDS in 10 mM Tris · HCl pH 8.2. Finally, the protein A-sepharose beads were suspended in 50 μ l sample buffer and heated for 5 min at 100°C.

(j) Chemical synthesis of deoxyoligonucleotides

Deoxyoligonucleotides were synthesized by the phosphotriester method in solution (De Rooy et al., 1979). For the synthesis of the *HpaII*-*Bam*HI linker an approach was chosen using protected dinucleotides as intermediates. As a consequence, the final step was a coupling of dCCGG with dATCCGG. The tetradecadeoxynucleotide dCATAGAATTCTATG was prepared by coupling dCATAG with dAATTCTATG. The latter nonamer was obtained from dAATT and dCTATG.

RESULTS

(a) Construction and identification of clones containing thaumatin cDNA sequences

The extensive purification of thaumatin-specific mRNA was found to be a prerequisite since cloning experiments using oligo(dT) cellulose-selected RNA still resulted in a significant amount of clones containing copied rRNA. To facilitate RNA analysis and isolation of mRNA encoding thaumatin, RNA isolated from arils of *Thaumatococcus* fruits was radioactively labelled using ATP:RNA adenylyl transferase and [α - ^{32} P]ATP (Sippel, 1973). The labelled RNA was purified by oligo(dT) cellulose affinity chromatography and analyzed by electrophoresis through polyacrylamide gels. Autoradiography of these gels showed the presence of a distinct band which was detectable only after a second purification of the RNA over oligo(dT) cellulose (Fig. 3). From this ob-

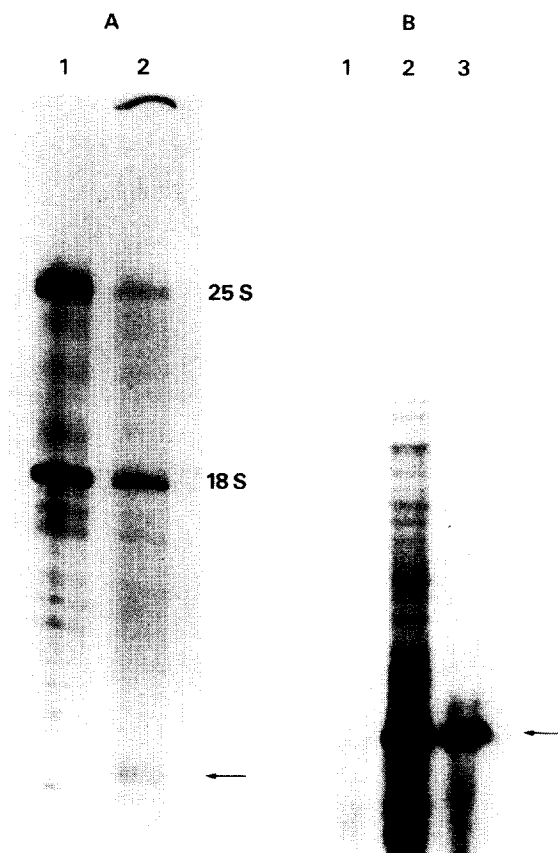


Fig. 3. (A) Autoradiogram of ^{32} P-labelled RNA extracted from *Thaumatococcus* arils purified once (lane 1) or twice (lane 2) on an oligo(dT) cellulose column. The positions of the plant ribosomal RNAs are indicated by 25S and 18S and the thaumatin-encoding mRNA by an arrow. (B) Fluorogram of [^{35} S]methionine-labelled proteins synthesized in vitro under the direction of either total polyadenylated RNA extracted from *Thaumatococcus* arils (lane 2) or the mRNA species eluted from the polyacrylamide gel (lane 3). In lane 1 no RNA was added. The position of the in vitro-synthesized thaumatin is indicated by an arrow.

servation it was concluded that the band represents an abundantly occurring, poly(A)-containing mRNA species. According to its electrophoretic mobility this mRNA molecule was estimated to be about 1000 nucleotides long. Elution of the particular mRNA species from the polyacrylamide gel followed by translation in a wheat germ-protein synthesizing system showed that the mRNA preparation mainly directed the synthesis of thaumatin (Fig. 3).

About 0.5 μ g of the highly purified, thaumatin-encoding mRNA material was used for single-

stranded cDNA synthesis using oligo(dT)₁₀ as a primer for AMV reverse transcriptase. The resulting cDNA was rendered double-stranded by *E. coli* DNA polymerase I-large fragment. After nuclease S1 digestion, the double-stranded DNA was subjected to velocity sedimentation through a glycerol gradient and fractions containing DNA fragments longer than 750 nucleotide pairs were collected. These fragments were tailed with deoxycytidine residues and annealed into the dG-tailed *Pst*I site of plasmid pBR322 DNA. The hybrid molecules were used to transform *E. coli* K-12, strain HB101, yielding 106 tetracycline-resistant, ampicillin-sensitive transformants.

Plasmid DNA isolated from 10 transformants was bound to diazotized (DBM) paper and hybridized to polyadenylated RNA isolated from arils. After washing the DBM paper, the hybridized mRNA species were isolated and translated in a wheat germ-protein synthesizing system.

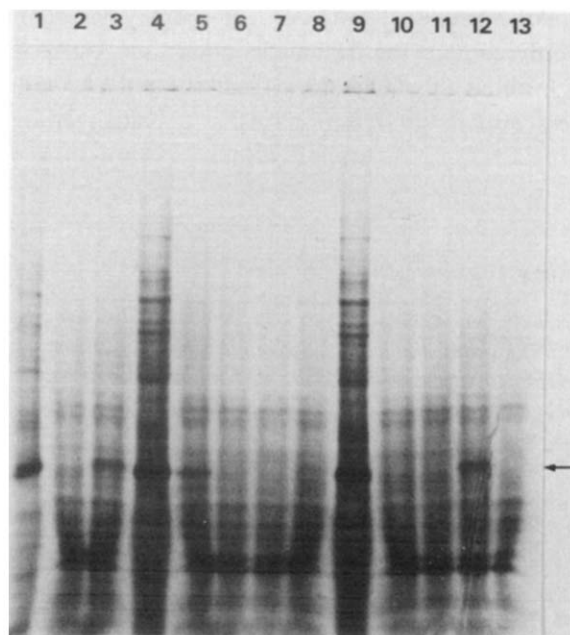


Fig. 4. Identification of recombinant plasmids containing cDNA sequences encoding thaumatin. Fluorogram of [³⁵S]cysteine-labelled proteins synthesized under the direction of either selected mRNA species (lanes 2, 3, 5, 6, 7, 8, 10, 11, 12, 13) or total polyadenylated *Thaumatooccus* RNA (lanes 1, 4, 9). The position of in vitro synthesized thaumatin is indicated by an arrow. The presence of thaumatin among the translation products shown in lanes 3, 5 and 12 is evident.

Three of these RNA preparations directed the synthesis of a protein which, upon analysis by SDS-gel electrophoresis, was shown to co-migrate with thaumatin (Fig. 4 lanes 3,5,12). Nucleotide sequence analysis of parts of the *Pst*I-inserts of the 3 plasmid DNA molecules unambiguously proved the identity of the clones. According to these data each of the three individual plasmid DNA molecules contained double-stranded DNA inserts whose nucleotide sequences could be aligned with the known amino acid sequence of thaumatin. When finally all 106 transformants were screened by a modified Grunstein-Hogness in situ colony hybridization procedure (Thayer, 1979), 21 were found to contain thaumatin-encoding DNA sequences.

(b) The nucleotide sequence of the thaumatin mRNA

From the 21 thaumatin-encoding transformants two plasmids, pUR100 and pUR110, were chosen and analysed in detail. The entire *Pst*I inserts of these plasmids were sequenced by both the chemical procedure of Maxam and Gilbert (1980) and the enzymatic dideoxy/nick-translation procedure of Maat and Smith (1978). Whereas neither insert covers the complete thaumatin mRNA, the 5'-end of the RNA molecule was sequenced indirectly by the primed synthesis/dideoxy procedure (Fig. 5). For the purpose of priming, a double-stranded DNA fragment (positions 36–105) isolated from pUR100 was treated with exonuclease III. This yielded a single-stranded primer, favouring DNA/RNA hybridization under the conditions used. There is still some ambiguity about the exact structure of the 5'-terminus since the primed synthesis/dideoxy procedure does not allow a definitive conclusion as to whether the nucleotide sequence immediately downstream of the presumptive cap structure contains two or three A's. The nucleotide sequence of the thaumatin mRNA is given in Fig. 6, together with the location of the areas covered by the respective *Pst*I inserts of pUR100 and pUR110.

By arranging the termination triplets in three phases a large translational reading frame is found with an ATG as putative initiation triplet at position 32 and TAA as termination triplet at position



Fig. 5. Autoradiogram of a polyacrylamide slab gel showing the 5'-terminal nucleotide sequence of the thaumatin mRNA as determined by the primed synthesis/dideoxy procedure.

736. The amino acid sequence of thaumatin I, as determined by Iyengar et al. (1979) via direct protein sequencing procedures, can be matched with the nucleotide sequence between positions 98 and 718. Comparison of this amino acid sequence

with the one deduced from the cloned cDNA reveals several discrepancies (Table I). They reflect differences between thaumatin forms I and II as became apparent from a comparison of tryptic peptide maps and amino acid composition analysis (P.D. van Wassenaar and H. van der Wel, personal communication).

According to the amino acid sequence studies no difference exists in the length of the proteins thaumatin I and II. The primary translation product (preprothaumatin) of the cloned cDNA is therefore longer than the mature thaumatin II molecule. This, quite surprisingly, implies post-transcriptional processing at both the amino and the carboxy terminus of the protein. The amino terminal extension of 22 amino acids in length is hydrophobic and closely resembles a signal sequence. According to the signal hypothesis (Blobel and Dobberstein, 1975) such a polypeptide sequence is involved in the process of cotranslational secretion. The carboxyterminal extension of six amino acids is very acidic in character, in contrast to the very basic mature thaumatin.

Polyadenylated mRNAs from eukaryotes commonly contain the hexanucleotide (5') AAUAAA (3'), about 20 nucleotides from the poly(A) tail. This apparently functions as a polyadenylation

TABLE I

Sequence divergencies between the thaumatin II-protein sequence deduced from the cloned cDNA and the protein sequence of thaumatin I derived by Iyengar et al. (1979)

Presumptive codons for the amino acids in the latter sequence have been underlined.

Number of aa	Thaumatin II		Thaumatin I		
	codon	aa	codon	aa	
46	AAG	lys	<u>AAC</u> , <u>AAT</u>	asN	
63	CGC	arg	<u>AGC</u> , <u>AGT</u> , <u>TCC</u> , <u>TCG</u> , <u>TCT</u> , <u>TCA</u>	ser	
67	CGG	arg	<u>AAG</u> , <u>AAA</u>	lys	
76	CAG	gln	<u>CGT</u> , <u>CGC</u> , <u>CGG</u> , <u>AGG</u> , <u>CGA</u> , <u>AGA</u>	arg	
113	GAC	asp	<u>AAC</u> , <u>AAT</u>	asN	

S10

(5') AAAG CGC AGC CTC AAT TGG CAT CAT ACA TCA  met ala ala thr thr cys phe phe phe leu

S1

S20 S22 1 50

phe pro phe leu leu leu leu thr leu ser arg ala ALA THR PHE GLU ILE VAL ASN ARG
TTC CCC TTC CTC CTC CTC CTC ACG CTC TCC CGC GCT GCC ACC TTC GAG ATC GTC AAC CGC

10 20 100

CYS SER TYR THR VAL TRP ALA ALA ALA SER LYS GLY ASP ALA ALA LEU ASP ALA GLY GLY
TGC TCC TAC ACC GTG TGG GCG GCC GCC TCC AAA GGC GAC GCC GCC CTG GAC GCC GGC GGC

30 150 40

ARG GLN LEU ASN SER GLY GLU SER TRP THR ILE ASN VAL GLU PRO GLY THR LYS GLY GLY
CGC CAG CTC AAC TCG GGA GAG TCC TGG ACC ATC AAC GTA GAA CCC GGC ACC AAG GGT GGC

50 200 60

LYS ILE TRP ALA ARG THR ASP CYS TYR PHE ASP ASP SER GLY ARG GLY ILE CYS ARG THR
AAA ATC TGG GCC CGC ACC GAC TGC TAT TTC GAC GAC AGC GGC CGC GGC ATC TGC CGC ACC

70 250 80 300

GLY ASP CYS GLY GLY LEU LEU GLN CYS LYS ARG PHE GLY ARG PRO PRO THR THR LEU ALA
GGC GAC TGC GGC GGC CTC CTC CAG TGC AAG CGC TTC GGC CGG CCG CCC ACC ACG CTG GCG

90 100 350

GLU PHE SER LEU ASN GLN TYR GLY LYS ASP TYR ILE ASP ILE SER ASN ILE LYS GLY PHE
GAG TTC TCG CTC AAC CAG TAC GGC AAG GAC TAC ATC GAC ATC TCC AAC ATC AAA GGC TTC

110 120 400

ASN VAL PRO MET ASP PHE SER PRO THR THR ARG GLY CYS ARG GLY VAL ARG CYS ALA ALA
AAC GTG CCG ATG GAC TTC AGC CCG ACC ACG CGC GGC TGC CGC GGG GTG CGG TGC GCC GCC

130 140 450

ASP ILE VAL GLY GLN CYS PRO ALA LYS LEU LYS ALA PRO GLY GLY GLY CYS ASN ASP ALA
GAC ATC GTG GGG CAG TGC CCG GCG AAG CTG AAG GCG CCG GGG GGT GGT TGC AAC GAT GCG

150 500 160

CYS THR VAL PHE GLN THR SER GLU TYR CYS CYS THR THR GLY LYS CYS GLY PRO THR GLU
TGC ACC GTG TTC CAG ACG AGC GAG TAC TGC TGC ACC ACG GGG AAG TGC GGG CCG ACG GAG

170 550 180 600

TYR SER ARG PHE PHE LYS ARG LEU CYS PRO ASP ALA PHE SER TYR VAL LEU ASP LYS PRO
TAC TCG CGC TTC TTC AAG AGG CTT TGC CCG GAC GCG TTC AGT TAT GTC CTG GAC AAG CCA

190 200 650 207

THR THR VAL THR CYS PRO GLY SER SER ASN TYR ARG VAL THR PHE CYS PRO THR ALA leu
ACC ACC GTC ACC TGC CCC GGC AGC TCC AAC TAC AGG GTC ACT TTC TGC CCT ACT GCC CTT

700

glu leu glu asp glu ***
GAA CTT GAA GAC GAG TAA GAG GAT GAA GAC GGA CAC TGA GGA TAC GCA ATA AAA GAA TAA

750

GAT GAT AAG CAA TTA TAT CAA TAA AAA GGG TAC GTG GTT TAC GTC GAG AAG GCA TCA GCT

800

TGG GAG GAA AAG TGT TAT AAA ATA TGT GTG TGG GGA TTG GCT AAT TAA ACT TGT AAA ATA

850 900

TAA ATA AAA GTT CTC CGT TTT GAG GTG TGC(A)_n (3')

Fig. 6. DNA sequence corresponding to the thaumatin II mRNA. The presumptive amino acid sequence of the preprothaumatin is given. The S-numbering denotes the amino acids in the putative signal sequence. "Polyadenylation signals" dAATAAA are dottedly underlined. The sequence in between the triangles was contained in the *Pst*I-insert of pUR100 (pos. 31-795); the sequence to the right of the circle including the poly(A) tail was contained in the *Pst*I-insert of pUR110.

signal. Inspection of the sequence, displayed in Fig. 6, reveals that this stretch occurs three times in the 3' untranslated region, one of which is present at the expected position.

(c) Expression of preprothaumatin in *E. coli*

To express preprothaumatin in *E. coli* two "expression plasmids" were constructed: one containing two *lacUV5*-regulons (pUR201) and another containing the *E. coli* tryptophan regulon (pUR301). The experimental plan for the incorporation of a modified preprothaumatin-encoding DNA fragment into pUR201 is outlined in Fig. 2. The removal of the dG/dC stretch, which is present upstream of the methionine initiation triplet of preprothaumatin in the insert of pUR100, is a prerequisite for the functioning of the bacterial ribosome-binding site in expression experiments. Selective elimination of the dG/dC stretch proved impossible enzymatically since this region seemed inert to the 3' → 5' exonucleolytic activity of either *E. coli* or T4 DNA polymerase. Therefore the more elaborate route shown in Fig. 2 was developed.

In addition to the endproduct pUR521, a plasmid was constructed which contained the

tryptophan regulon (pUR531). To this end, the *EcoRI*-insert of pUR521 was inserted into the *EcoRI* site of pUR301. Plasmid pUR521 also formed the basic material for the construction of plasmids in which the thaumatin insert covered the complete mRNA downstream of the ATG initiation codon. For this purpose inserts from pUR521 and pUR110 were linked together via the *RruI*-site at position 600 (construction not shown). Integration of this insert into pUR201 and pUR301 resulted in plasmids pUR528 (*lac*) and pUR538 (*trp*), respectively. Fig. 7 displays a schematic presentation of pUR538.

The plasmids pUR521, pUR528, pUR531 and pUR538 served to direct the synthesis of the plant protein preprothaumatin in *E. coli*. This was monitored by SDS-polyacrylamide gel electrophoresis of immunoprecipitated proteins obtained from cells pulse-labeled with [³⁵S]cysteine. As shown in Fig. 8, one of the proteins immunoprecipitated from cells harbouring plasmids pUR531 and pUR538 co-migrated with wheat germ-synthesized preprothaumatin. Such a protein is absent in extracts of

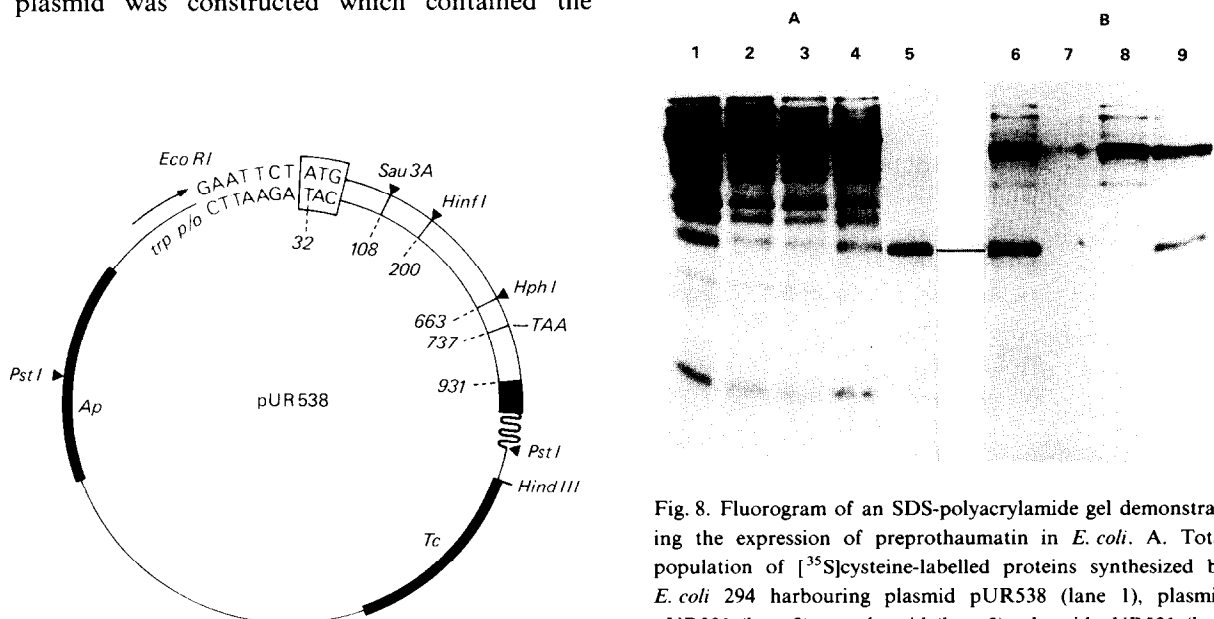


Fig. 7. Schematic presentation of plasmid pUR538, carrying a full-length cDNA copy of thaumatin II mRNA downstream of the ATG-initiation codon. Single cleavage sites for *Sau3A*, *HinfI* and *HphI* in the thaumatin gene are indicated. ■, poly(dA/dT) stretch; wavy line, dG/dC-stretch. Numbers refer to the DNA sequence given in Fig. 6.

Fig. 8. Fluorogram of an SDS-polyacrylamide gel demonstrating the expression of preprothaumatin in *E. coli*. A. Total population of [³⁵S]cysteine-labelled proteins synthesized by *E. coli* 294 harbouring plasmid pUR538 (lane 1), plasmid pUR301 (lane 2), no plasmid (lane 3), plasmid pUR531 (lane 4). B. Immunoprecipitated [³⁵S]cysteine-labelled proteins synthesized by *E. coli* 294 harbouring plasmid pUR531 (lane 6), no plasmid (lane 7), plasmid pUR301 (lane 8) and plasmid pUR538 (lane 9). The protein in lane 5 represents immunoprecipitated pre-prothaumatin synthesized in vitro in a wheat germ protein-synthesizing system.

plasmid-free cells and cells carrying plasmid pUR301. On the basis of these observations one must conclude that the protein synthesized is preprothaumatin. Besides the latter protein some high- M_r material is present in the immunoprecipitates. It is not known whether this is due to aspecific absorption or to a weak anti *E. coli* activity of the rabbit antiserum used. Direct analysis of the pulse-labeled cells also shows the synthesis of preprothaumatin (Fig. 8A). This is indicated by the presence of an extra band in the lanes with the extracts from cultures harbouring pUR531 and pUR538. Synthesis of preprothaumatin under control of the double-*lacUV5*-regulon was demonstrated analogously using the plasmids pUR521 and pUR528. Comparison of the amounts of preprothaumatin synthesized under either *trp* or *lac* regulon control revealed no perceptible difference. Procedures for detection of plasmid-coded proteins such as the chloramphenicol amplification procedure of Neidhardt et al. (1980) and in vitro translation of plasmid DNA in a coupled transcription-translation system likewise demonstrated the synthesis of the preprothaumatin protein. It is noteworthy that, under these conditions, contaminants in the immunoprecipitates were absent.

DISCUSSION

We have demonstrated that a DNA copy of preprothaumatin-encoding mRNA is expressed in *E. coli* under transcriptional control of either the *lac* regulon or the *trp* regulon. Comparison of the amino acid sequence as deduced from the cloned DNA with the amino acid sequence of thaumatin I and comparison of the fingerprints of tryptic digests of the various thaumatins indicates that the mRNA-derived cDNA of preprothaumatin II has been cloned. Its sequence displays several discrepancies with the thaumatin I sequence (Table I). Four out of five of the amino acid substitutions can be pinpointed as single base changes in the mRNA molecule. The conversion of arginine to lysine (position 67), however, requires the substitution of two nucleotides (CGG → AAG). Surprisingly, it is a basic to basic amino acid change, which might indicate the importance of a positive

charge at that position in the thaumatin molecule. Although processing of a primary translation product resulting in the removal of a hydrophobic amino-terminal peptide is commonly observed, processing removing an acidic carboxy-terminal peptide is not. On the basis of our nucleotide sequence data and the data obtained by carboxypeptidase digestion of thaumatin II, we must conclude that some time after translation six carboxy-terminal amino acids are cleaved off from the protein molecule. Since, to our knowledge, a similar processing step has never been described, it is difficult to ascribe a biological function to this phenomenon. A parallel, however, may be drawn with observations made by Highfield and Ellis (1978) in studies on the compartmentalization of the small subunit of chloroplast RuBPCase. According to their envelope-carrier hypothesis proteins destined to function in the chloroplast contain specific sites which are recognized by receptors in the chloroplast envelope. Removal of an acidic part of the RuBPCase causes a conformational change leading to the transport of the protein across the chloroplast membrane. Since preliminary experiments suggest that thaumatin is stored in the aril in vesicle-like organelles, one might speculate on the basis of the above mentioned hypothesis, that the acidic carboxy-terminal extension of the — extremely basic — thaumatin molecule serves as a signal for compartmentalization.

Currently we are constructing plasmids which will permit the synthesis of the mature forms of thaumatin. Physiological studies are being undertaken to assess the sweet character of the bacterially produced preprothaumatin and its mature form. The molecular cloning and expression of the thaumatin II structural gene in *E. coli* may be useful for obtaining a better insight into the structural requirements of thaumatin to elicit its sweet taste. Specific gene alterations caused by site-directed mutagenesis will be an important tool in this study. Moreover, this plant structural gene is an excellent probe to test gene transfer with the *Agrobacterium* Ti-plasmid vector system. Experiments of this kind are in progress, in collaboration with Drs. J. Schell and M. Van Montagu of the University of Ghent, Belgium.

ACKNOWLEDGEMENTS

We thank Drs. M. van Montagu and J. Schell (University of Ghent) for their hospitality and the use of their laboratory in the cloning experiments. Drs. A. de Waard and R. Hallewell are acknowledged for supplying *RruI* and plasmid *ptrp* ED5-1, respectively. Miss W. Bos supplied excellent technical assistance. Thanks are also due to Dr. W. Hesselink (CLB, Amsterdam) for helpful advice and to H. van der Wel, D. van Wassenaar and Dr. F. van der Ouderaa for supplying various materials and for helpful discussions.

REFERENCES

- Birnboim, H.C. and Doly, J.: A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucl. Acids Res.* 7 (1979) 1513–1523.
- Blobel, G. and Dobberstein, D.: Transfer of protein across membranes. *I. J. Cell. Biol.* 67 (1975) 835–851.
- Buell, G.N., Wickens, M.P., Paguar, F. and Schimke, R.T.: Synthesis of full length cDNAs from four partially purified oviduct mRNAs. *J. Biol. Chem.* 253 (1978) 2471–2482.
- Davies, J.W. and Kaesberg, P.: Translation of virus mRNA: Synthesis of bacteriophage QB proteins in a cell-free extract from wheat embryo. *J. Virol.* 12 (1973) 1434–1441.
- De Rooij, J.F.M., Wille-Hazeleger, G., Van Deursen, P.H., Serdijn, J. and Van Boom, J.H.: Synthesis of complementary DNA fragments via phosphotriester intermediates. *Recl. Trav. Chim. Pays-Bas* 98 (1979) 537–548.
- Gatenby, A.A., Castleton, J.A. and Saul, M.W.: Expression in *Escherichia coli* of maize and wheat chloroplast genes for the large subunit of ribulose 1,5-biphosphate carboxylase. *Nature* 291 (1981) 117–121.
- Glaser, D., Hellekant, G., Brouwer, J.N. and Van der Wel, H.: The taste responses in primates to the proteins thaumatin, monellin and their phylogenetic implications. *Folia Primatol.* 29 (1978) 56–63.
- Goeddel, D.V., Heyneker, H.L., Hozumi, T., Arentzen, R., Itakura, K., Yansura, D.G., Ross, M.J., Miozarri, G., Crea, R. and Seeburg, P.: Direct expression in *Escherichia coli* of a DNA sequence coding for human growth hormone. *Nature* 281 (1979) 544–548.
- Hallewell, R.A. and Emtage, S.: Plasmid vectors containing the tryptophan operon promoter suitable for efficient regulated expression of foreign genes. *Gene* 9 (1980) 27–47.
- Hamlyn, P.H., Brownlee, G.G., Chang, C.C., Gait, M.J. and Milstein, C.: Complete sequence of constant and 3' noncoding regions of an immuno globulin mRNA using the dideoxy method of RNA sequencing. *Cell* 15 (1978) 1067–1075.
- Higginbotham, J.D. and Hough, C.A.M.: Useful taste properties of amino acids and proteins, in Birch, G.C., Brennan, J.G. and Parker, K.J. (Eds.). *Sensory Properties of Foods*, Applied Sciences, London, 1977, pp. 129–149.
- Highfield, P.E. and Ellis, R.J.: Synthesis and transport of the small subunit of chloroplast ribulose biphosphate carboxylase. *Nature* 271 (1978) 420–424.
- Iyengar, R.B., Smits, P., Van der Wel, H., Van der Ouderaa, F.J.G., Van Brouwershaven, J.H., Ravenstein, P., Richters, G. and Van Wassenaar, P.D.: The complete amino acid sequence of the sweet protein thaumatin I. *Eur. J. Biochem.* 96 (1979) 193–204.
- Maat, J. and Smith, A.J.H.: A method for sequencing restriction fragments with dideoxynucleotide triphosphate. *Nucl. Acids Res.* 5 (1978) 4537–4545.
- Maxam, A.M. and Gilbert, W.: Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* 65 (1980) 499–560.
- Murray, N.E., Bruce, S.A. and Murray, K.: Molecular cloning of the DNA ligase gene from bacteriophage T4; II: Amplification and preparation of the gene product. *J. Mol. Biol.* 132 (1979) 493–505.
- Neidhart, F.C., Wirth, R., Smith, M.W. and Van Bogelen, R.: Selective synthesis of plasmid-coded proteins by *Escherichia coli* during recovery from chloramphenicol treatment. *J. Bacteriol.* 143 (1980) 535–537.
- Sippel, A.E.: Purification and characterization of ATP:RNA adenylyl transferase from *Escherichia coli*. *Eur. J. Biochem.* 37 (1973) 31–40.
- Squires, C., Lee, F., Bertrand, K., Squires, C.L., Bronson, M.J. and Yanofsky, C.: Nucleotide sequence of the 5'-end of tryptophan messenger RNA of *Escherichia coli*. *J. Mol. Biol.* 103 (1976) 351–381.
- Sutcliffe, J.G.: Complete nucleotide sequence of the *Escherichia coli* plasmid pBR 322. *Cold Spring Harbor Symp. Quant. Biol.* 43 (1979) 77–90.
- Thayer, R.E.: An improved method for detecting foreign DNA in plasmids of *Escherichia coli*. *Anal. Biochem.* 98 (1979) 60–63.
- Van der Wel, H. and Bel, W.J.: Effect of acetylation and methylation on the sweetness intensity of thaumatin I. *Chem. Senses Flavor* 2 (1976) 211–218.
- Van der Wel, H. and Loeve, K.: Isolation and characterisation of thaumatin I and II, the sweet tasting proteins from *Thaumatococcus daniellii* Benth. *Eur. J. Biochem.* 31 (1972) 221–225.
- Wickens, M.P., Buell, G.N. and Schimke, R.T.: Synthesis of double stranded DNA complementary to lysozyme, ovomucoid and ovalbumin mRNAs. *J. Biol. Chem.* 253 (1978) 2483–2495.
- Williams, J.G., Lloyd, M.M. and Devine, J.M.: Characterization and transcription analysis of a cloned sequence derived from a major developmentally regulated mRNA of *D. discoideum*. *Cell* 17 (1979) 903–913.
- Wyckoff, M., Rodbard, D. and Chrambach, A.: Polyacrylamide gel electrophoresis in SDS-containing buffers using multiphase buffer systems. *Anal. Biochem.* 78 (1977) 459–482.
- Yanofsky, C.: Attenuation in the control of expression of bacterial operons. *Nature* 289 (1981) 751–758.

Communicated by A.J. van der Eb.