The EMBO Journal vol.3 no.4 pp.835–846, 1984

The complete nucleotide sequence of the TL-DNA of the *Agrobacterium tumefaciens* plasmid pTiAch5

J. Gielen, M. De Beuckeleer[1], J. Seurinck[1,4], F. Deboeck[2], H. De Greve[2], M. Lemmers[1], M. Van Montagu[1,2] and J. Schell[1,3]*

Histologisch Instituut and [1]Laboratorium voor Genetica, Rijksuniversiteit Gent, B-9000 Gent, [2]Laboratorium voor Genetische Virologie, Vrije Universiteit Brussel, B-1640 St.-Genesius-Rode, Belgium, [3]Max-Planck-Institut für Züchtungsforschung, D-5000 Köln 30, FRG
[4]Present address: Plant Genetic Systems, B-9000 Gent, Belgium

*To whom reprint requests should be sent
Communicated by J. Schell

We have determined the complete primary structure (13 637 bp) of the TL-region of *Agrobacterium tumefaciens* octopine plasmid pTiAch5. This sequence comprises two small direct repeats which flank the TL-region at each extremity and are involved in the transfer and/or integration of this DNA segment in plants. TL-DNA specifies eight open-reading frames corresponding to experimentally identified transcripts in crown gall tumor tissue. The eight coding regions are not interrupted by intervening sequences and are separated from each other by AT-rich regions. Potential transcriptional control signals upstream of the 5′ and 3′ ends of all the transcribed regions resemble typical eukaryotic signals: (i) transcriptional initiation signals ('TATA' or Goldberg-Hogness box) are present upstream to the presumed translational start codons; (ii) 'CCAAT' sequences are present upstream of the proposed 'TATA' box; (iii) polyadenylation signals are present in the 3′-untranslated regions. Furthermore, no Shine-Dalgarno sequences are present upstream of the presumed translational start codons.
Key words: Agrobacterium tumefaciens/T-DNA/nucleotide sequence

Introduction

One of the remarkable properties of the Ti plasmids of *Agrobacterium* is their natural capacity to transfer, insert, and express a particular DNA segment of the Ti plasmid in plant cells (for recent reviews, see Nester and Kosuge, 1981; Bevan and Chilton, 1982; Caplan *et al.*, 1983; Zambryski *et al.*, 1983). Depending on the host plant and on the nature of Ti plasmid present in the inciting *Agrobacterium* strain, the transformation event results in crown gall or hairy-root or woolly-knot disease (see Kahl and Schell, 1982).

The segment of Ti plasmid DNA which becomes stably inserted in the plant genome is called T-DNA (Chilton *et al.*, 1977; Lemmers *et al.*, 1980; Thomashow *et al.*, 1980). On the Ti plasmid this DNA segment is bordered by two direct-repeat sequences of 25 bp (Zambryski *et al.*, 1982, 1983; Yadav *et al.*, 1982; Holsters *et al.*, 1983). In the case of the octopine Ti plasmids, two regions of the Ti plasmid, called TL (T-left) and TR (T-right) (Thomashow *et al.*, 1980) according to their position on the standard octopine Ti plasmid map (De Vos *et al.*, 1981) can be transferred and inserted independently into the plant genome. The TL-DNA has been studied more extensively because it encodes essential functions involved in the neoplastic transformation of plant cells (De Beuckeleer *et al.*, 1981; Garfinkel *et al.*, 1981; Leemans *et al.*, 1982; Willmitzer *et al.*, 1982). The TL-DNA also comprises the functions found in common between octopine-type and nopaline-type Ti plasmids' T-regions (Depicker *et al.*, 1978; Chilton *et al.*, 1978; Engler *et al.*, 1981; Willmitzer *et al.*, 1983).

Recently, the nucleotide sequence of the octopine synthase gene (De Greve *et al.*, 1982a), of the gene for 'transcript 7' (Dhaese *et al.*, 1983), and of the gene for 'transcript 4' (Heidekamp *et al.*, 1983) were determined. Here we present the complete nucleotide sequence of the TL-DNA of the *Agrobacterium tumefaciens* plasmid pTiAch5.

Results and Discussion

Sequence determination

To determine the complete sequence of the octopine TL-region, different plasmids containing subfragments of the TL-DNA were constructed (Table I) from clones pGV0153 and pGV0201 (De Vos *et al.*, 1981) containing fragments *Bam*HI-8 and *Hind*III-1 (Figure 1), which overlap the complete TL-DNA region. Detailed physical maps of these subclones were established to facilitate the nucleotide sequencing. Plasmid DNA was cleaved with a particular restriction enzyme, and the resulting fragments were ^{32}P end-labeled either at their 5′ termini with polynucleotide kinase or at their 3′ termini with the Klenow fragment of DNA polymerase I. After strand separation or secondary restriction to separate the labeled extremities, the sequence was determined by the limited chemical cleavage method of Maxam and Gilbert (1980). Both DNA strands were sequenced to avoid mistakes that could occur in regions with a distinct secondary structure or by incorrect reading and processing of the sequence information. In addition, as methylated bases (Ohmori *et al.*, 1978) can interfere with correct reading of the sequence, all *Eco*RII sites located in the TL-region were used for sequencing. Furthermore, care was taken that all restriction sites used to generate fragments were resequenced by using another fragment containing an alternative site. Figure 2 gives an overview of the sequence strategy.

Sequence analysis

An uninterrupted sequence of 13 637 bp including the whole TL-DNA of pTiAch5 was determined, and is displayed in the conventional orientation in Figure 3. The numbering starts at the *Hind*III site bordering fragments 14 and 18c, which is located 308 bp to the left of the left TL-DNA terminus sequence.

Termini sequences. The TL-region is flanked at both extremities (position 308 and 13 459) by direct repeats of 24 bases, which are believed to be important for the transfer of the TL-DNA segment (Zambryski *et al.*, 1982; Simpson *et al.*, 1982; Holsters *et al.*, 1983).

Table I. Bacterial strains and plasmids

	Antibiotic resistance	Characteristics	Origin
Bacterial strains			
K514		*thr leu thi hsd*R	Colson *et al.* (1965)
SK383	Sm	F⁻ Arg⁻ *his*4, Ilv⁻ *lac*MS286 φ80dII*lac*BK1 Sup⁻ *dam*4	S. Kurshner
Plasmids			
pGV0153	Ap	*Bam*HI-8 of pTiAch5 in pBR322	De Vos *et al.* (1981)
pGV117	Ap Cml	*Hind*III-18c of pTiAch5 in pBR325	Dhaese *et al.* (1983)
pGV714	Ap Cml	*Hind*III-22c of pTiAch5 in pBR325	This work
pGV715	Ap Cml	*Hind*III-36 of pTiAch5 in pBR325	This work
pGV716	Ap Cml	*Hind*III-*Bam*HI fragment overlapping the fragments *Bam*HI-8 and *Hind*III-1 in pBR325	This work
pGV0201	Ap	*Hind*III-1 of pTiAch5 in pBR325	De Vos *et al.* (1981)
pGV105	Ap Tc	*Eco*RI-19a of pTiAch5 in pBR325	De Greve *et al.* (1982a)
pGV99	Ap Clm	*Bam*HI-17a of pTiAch5 in pBR325	De Greve *et al.* (1982a)
pGV101	Ap Clm	*Bam*HI-17a of pTiAch5 in pBR325	This work
pGV100	Ap Clm	*Bam*HI-28 of pTiAch5 in pBR325	This work
pGV732	Ap Clm	*Ava*I deletion of pGV101	This work
pGV733	Ap Clm	*Bcl*I deletion of pGV732	This work
pGV734	Ap Clm	*Bcl*I deletion of pGV0201	This work

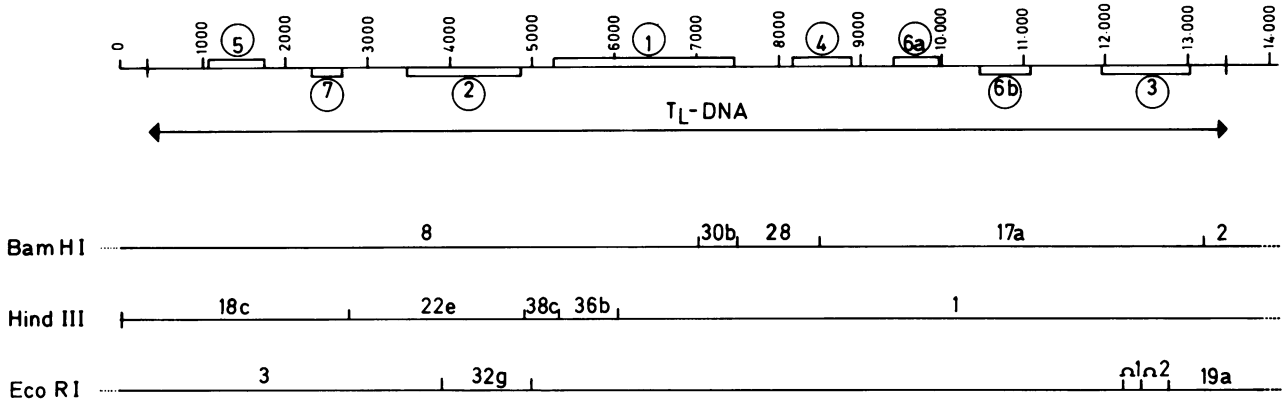

Fig. 1. Restriction map of the TL-DNA of the octopine Ti plasmid pTiAch 5. **Upper portion**: the position of the open-reading frames are presented by open boxes and numbered according to Willmitzer *et al.* (1982). The polarity of the open-reading frames is indicated as follows: open boxes above the line are transcribed from left to right and open boxes below the line are transcribed from right to left. The extent of the TL-DNA is indicated by an arrow and is delimited by the termini boxes (heavy vertical bars). **Lower portion**: a restriction map of the TL-DNA region is shown for the restriction enzymes *Bam*HI, *Hind*III, and *Eco*RI.

A computer search of the complete TL-region for DNA sequences displaying homologies with these direct repeats revealed 10 related DNA sequences. These sequences are listed in Table II. Genetic and physical data indicate that some of these sequences might also be used *in vivo* during transfer and integration of the TL-DNA. Firstly, the sequence (position 11 798) present in the 3'-untranslated region of the octopine synthase gene has been noted by Holsters *et al.* (1983). If this sequence is recognized as a left terminus sequence, the presence of the abbreviated T-DNA found in the octopine-positive regenerate plants rGV1 and rGV5 (De Greve *et al.*, 1982b) can be explained. Alternatively, if this sequence is recognized as a right terminus sequence, instead of the normal terminus sequence, tumor lines containing a shorter TL-DNA which do not synthesize octopine (Thomashow *et al.*, 1980; De Beuckeleer *et al.*, 1981; Ooms *et al.*, 1982) are formed. The origin of teratomas (unpublished results) expressing transcripts 4, 6a, 6b, octopine synthase, and possibly transcript 1, can be explained if the sequence (position 3750) located in transcript 2 is used as a left terminus sequence. Similarly, an abnormal plant (unpublished data) possibly containing transcript 4 and expressing the octopine synthase gene, could be explained if the sequence (position 7777) is used as a left terminus sequence. In addition, either the sequences at position 9078, 10 131, or 10 603 if used as a right terminus sequence, could explain the short TL-DNA observed in a *Petunia* tumor line P-Ach5 (De Beuckeleer *et al.*, 1981). Whether the other sequences also signalled the creation of abbreviated TL-DNAs is difficult to answer because in most cases the resulting transferred DNA

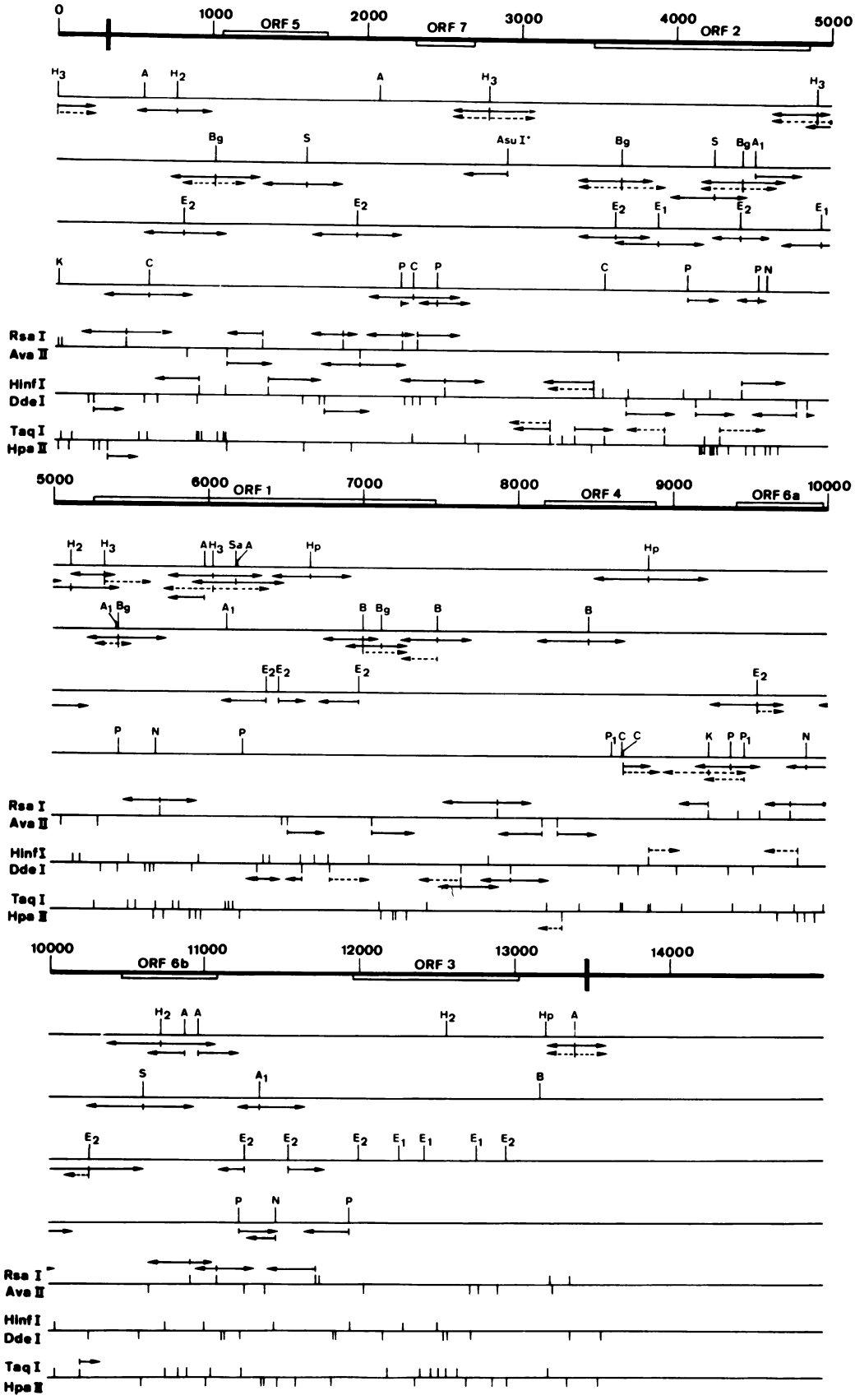

Fig. 2. Sequencing strategy. On a map of the TL-region of pTiAch5 the restriction sites for the following enzymes have been indicated: A, *Acc*I, A1, *Ava*I; Bg, *Bgl*II; C, *Cla*I; E1, *Eco*RI; E2, *Eco*RII; H2, *Hind*II; H3, *Hind*III; Hp, *Hpa*I; K, *Kpn*I; N, *Nae*I; P, *Pvu*II; P1, *Pst*I; S, *Sma*I; Sa, *Sal*I. The position and extent of each sequencing experiment is indicated by a full arrow for a 5′ to 3′ sequencing, and a dashed one for 3′ to 5′. Termini boxes are indicated by a heavy bar, and the open-reading frames corresponding to plant transcripts by open boxes. The polarity of the open-reading frames is indicated from left to right by drawing the open boxes above the line and from right to left by drawing the open boxes below the line.

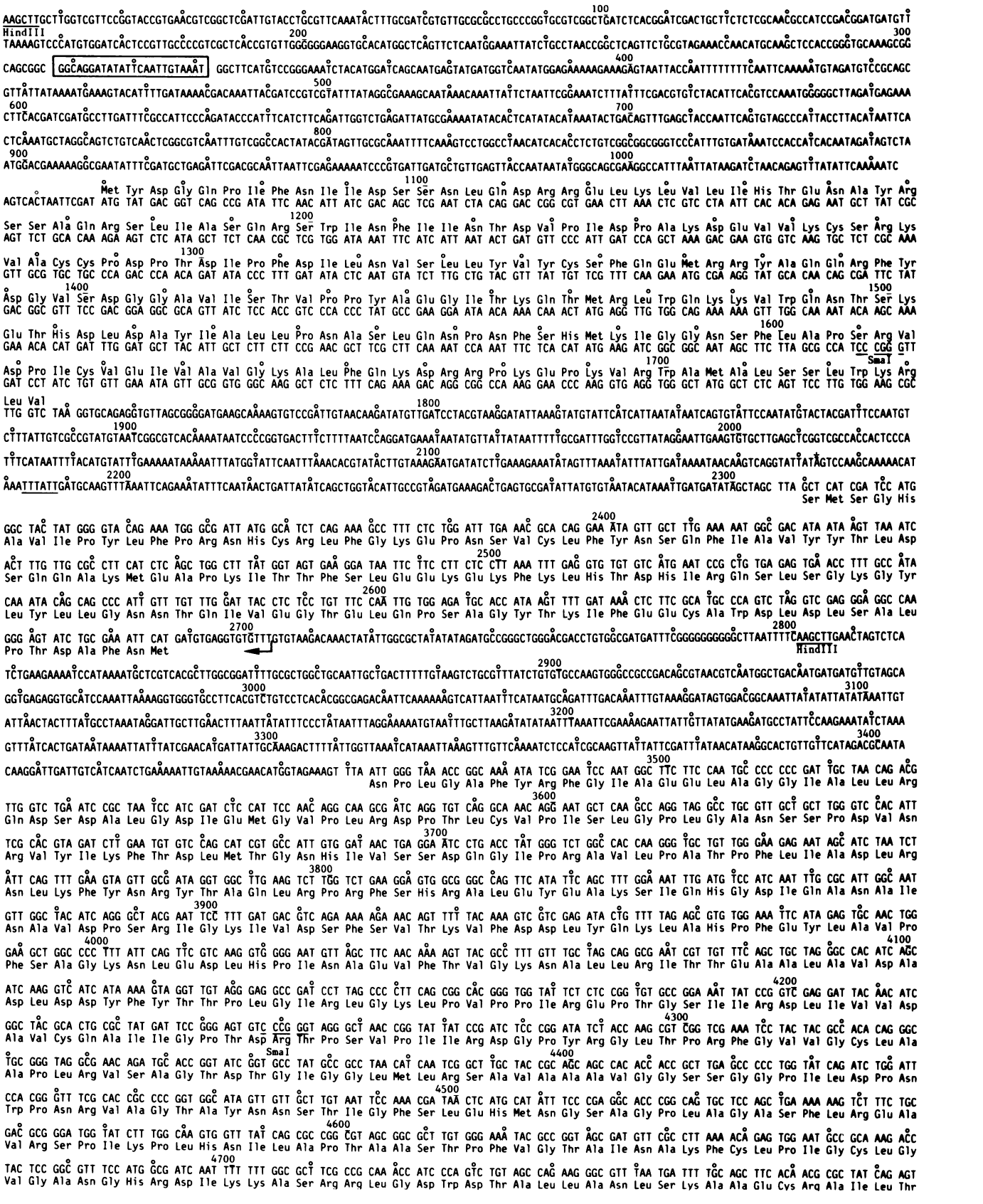

Figure 3(i)

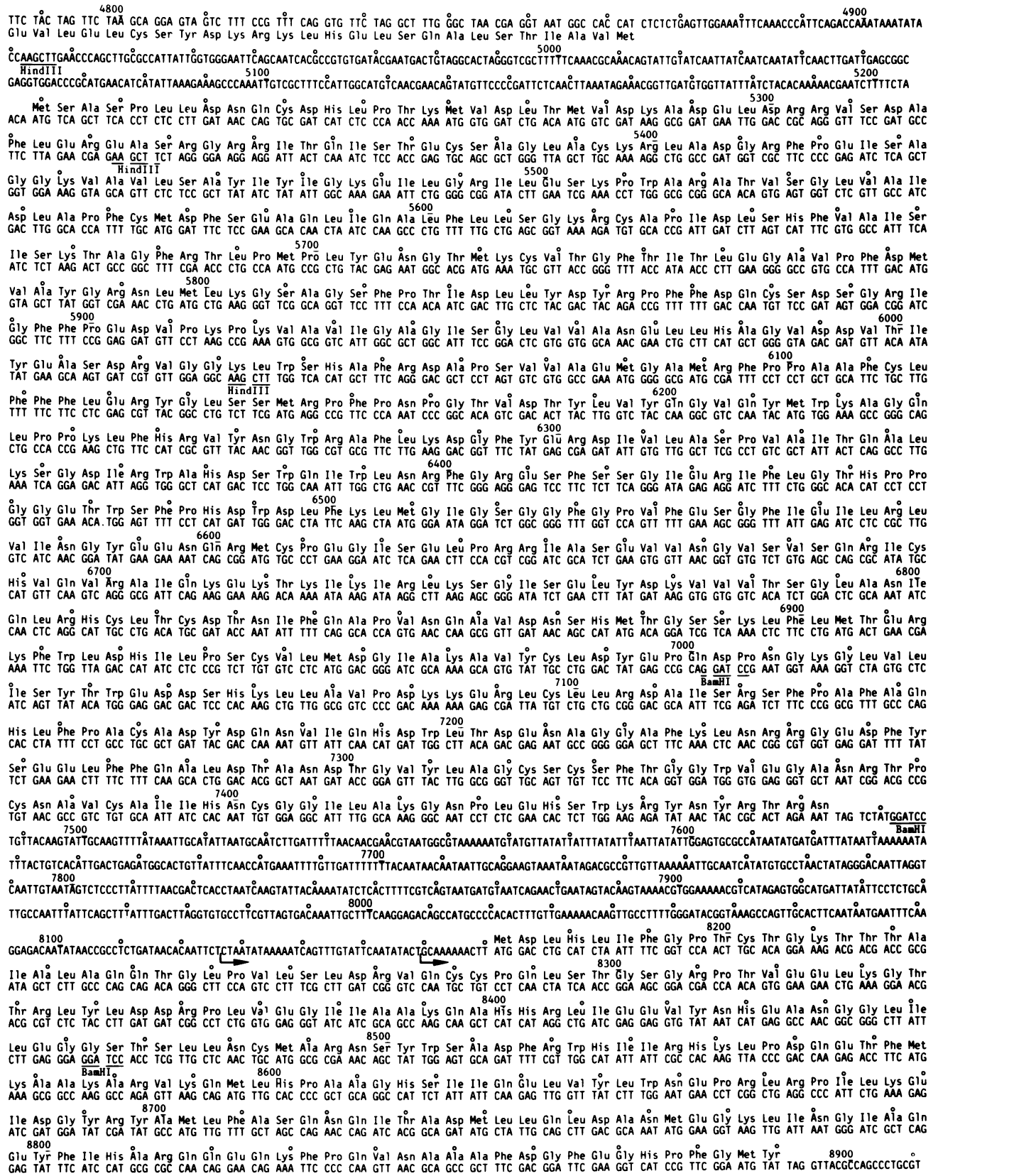

Figure 3(ii)

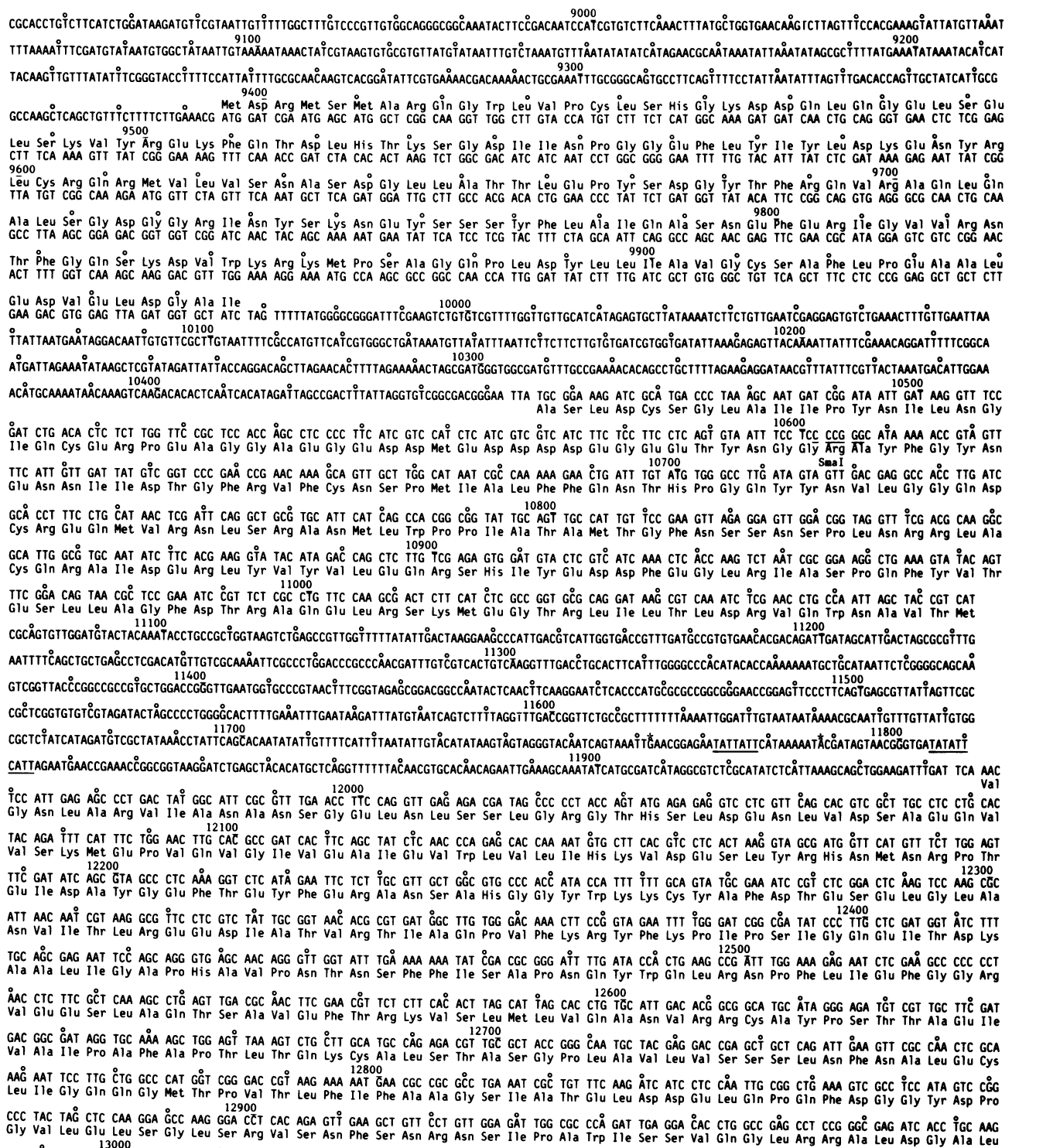

Figure 3(iii)

Fig. 3. Complete nucleotide sequence of the TL-region of pTiAch5. An uninterrupted sequence of 13 637 bp starting at the *Hind*III site bordering the fragments 14 and 18c covers the whole TL-region. The sequence is displayed in the conventional orientation along with the translation in amino acids for the coding sequences for which experimental evidence exists. The amino acid sequence is above the DNA sequence when transcription occurs from left to right, and below the sequence for the other orientation. The two direct repeats present at both extremities of the TL-DNA are indicated by a closed box. The mRNA start and the polyadenylated sites and signals of transcripts 3, 4 and 7 are indicated by an arrow. The polyadenylation signals of transcripts 3 and 7 are underlined and their polyadenylation sites are indicated by an asterix.

would not produce an easily detected altered phenotype in the transformed plant cells.

Size and position of coding sequences. The sequence between the 24-bp direct repeats was analyzed for possible translational open-reading frames. The 18 largest open-reading frames are presented in Table III. To evaluate which of these open-reading frames are actually used *in vivo*, their position was compared with the known positions of TL-DNA transcripts in octopine crown gall tissues (Willmitzer *et al.*, 1982). Seven of the open-reading frames did correspond with known transcripts. We tested whether or not some of the other open-reading frames might correspond to TL-DNA regions, whose transcripts might have gone undetected, by comparing their position with empty regions in the transcription map. This was the case only for open-reading frame m (Table III). Subsequently, a careful experimental analysis confirmed that this open-reading frame corresponded to an actual transcript (6b) (Willmitzer *et al.*, 1983; Joos *et al.*, 1983). The translation of these eight open-reading frames in amino acids is presented in Figure 3 and their codon usage is listed in Table IV. It was also tested whether open-reading frame p which is derived from the opposite strand of transcript 3 and which might code for a protein of 142 amino acids could correspond to an actual transcript. M13 mp2 phage DNA, containing the small *Eco*RI fragments Ω_1 and Ω_2 (Figure 1) located in the octopine synthase gene, were separately applied on nitrocellulose and hybridized with labeled mRNA isolated from tobacco crown gall tissues. Only the phage DNA spot containing the strand corresponding to transcript 3 (octopine synthase) hybridized with mRNA (data not shown).

We have applied the RNY algorithm described by Shepherd (1981) on the whole sequence of the TL-DNA (data not shown). Eight frames were detected and these correspond to the eight known transcribed regions.

The size and map position of several proteins, expressed by the T-DNA in transformed plant cells, or by the T-region in bacterial cell-free systems, have been recently determined (summarized in Table III). By hybridization selection and translation of T-DNA-encoded mRNA from octopine tumors, three proteins of 39, 27 and 14 kd were detected (Schröder and Schröder, 1982). The largest has been shown to

Table II. DNA sequences homologous to the 24-bp termini sequences

Left terminus sequence	GGCAGGATATATTCAATTGTAAAT	308 bp
	ACCAATTTTTTTTCAATTCAAAAA	407 bp
	CAGAGTTTATATTCAAAAATCAGT	1024 bp
	CCCAACAGATATACCCTTTGATAT	1293 bp
	CCTTTGATATACTCAATGTATCTT	1307 bp
	CATCTAATCTATTCAGTTTGAAGT	3750 bp
	GGGACAATTAGGTCAATTGTAATA	7777 bp
	TATAATGTGGCTATAATTGTAAAA	9078 bp
	TAAATGTTATATTTAATTCTTCTT	10 131 bp
	CCGGGCATAAAAACCGTAGTTTTC	10 603 bp
	CGGGTGATATATTCATTAGAATGA	11 798 bp
Right terminus sequence	GGCAGGATATATACCGTTGTAATT	13 459 bp

The TL-region sequence was compared with the left and the right terminus sequences using the comparison program written by Schroeder and Blattner (1982). All sequences sharing >50% homology with the terminus sequences were maintained.

Table III. Co-ordinates of open-reading frames of the TL-region DNA

Open region	Nucleotide First	Nucleotide Last	First ATG in frame	Σ AA	Mol. wt. Calculated (d)	Mol. wt. Observed (kd)	Correspondence
a	1054	1740	1060	226	25 635		Transcript 5
b	1569	1135	1512	125	14 310		
c	2726	2307	2687	126	14 219	14	Transcript 7
d	4124	4474	4232	80	8252		
e	4881	3460	4863	467	49 655	49	Transcript 2
f	5155	7476	5209	755	83 815	74	Transcript 1
g	6039	5659	5979	106	12 101		
h	6888	6622	6876	84	10 014		
i	7025	7513	7178	111	12 750		
j	8105	8893	8171	240	26 873	27	Transcript 4
k	8542	8294	8527	77	8858		
l	9344	9970	9395	191	21 335		Transcript 6a
m	11 160	10 453	11 076	207	23 320		Transcript 6b
n	11 142	11 405	11 178	75	8160		
o	11 581	11 092	11 353	86	9375		
p	12 020	12 460	12 032	142	16 455		
q	13 081	11 954	13 030	358	38 665	39	Transcript 3
r	13 203	12 901	13 203	100	11 331		

The table displays all the open-reading frames larger than 75 amino acids. The co-ordinates are those of the first nucleotide following the preceding stop, the last nucleotide of the stop codon and the A of the first ATG in frame. The length of the deduced protein (expressed in amino acids, ΣAA) and its mol. wt. has been calculated and is compared, when possible, with experimental data (Schröder and Schröder, 1982; Schröder *et al.*, 1981, 1983).

Table IV. Codon usage

		Transcripts										Transcripts										Transcripts										Transcripts							
		5	7	2	1	4	6a	6b	3			5	7	2	1	4	6a	6b	3			5	7	2	1	4	6a	6b	3			5	7	2	1	4	6a	6b	3
Phe	UUU	3	5	11	21	2	4	5	8	Ser	UCU	4	1	3	13	1	3	2	6	Tyr	UAU	6	6	8	12	7	7	4	5	Cys	UGU	2	0	3	8	1	3	1	0
	UUC	5	3	6	22	7	3	4	8		UCC	4	2	6	11	1	1	5	5		UAC	2	1	4	11	1	3	6	4		UGC	3	3	4	13	3	0	4	4
Leu	UUA	1	2	9	4	1	3	2	1		UCA	1	3	5	9	1	5	1	5	Stop	UAA	1	1	1	0	0	0	1	0	Stop	UGA	0	0	0	0	0	0	0	1
	UUG	5	3	7	13	6	4	4	7		UCG	4	1	3	6	2	2	0	4		UAG	0	0	0	1	1	1	0	0	Trp	UGG	5	1	2	14	3	2	2	4
	CUU	4	0	7	11	9	6	5	11	Pro	CCU	1	0	7	13	3	1	1	3	His	CAU	2	3	3	13	8	1	1	2	Arg	CGU	1	0	3	6	2	0	2	2
	CUC	5	3	6	16	2	3	1	9		CCC	4	2	8	3	4	1	0	3		CAC	1	1	6	4	2	1	1	3		CGC	4	1	7	6	2	1	4	3
	CUA	2	1	8	5	3	3	1	4		CCA	7	4	11	10	3	3	3	7	Gln	CAA	7	5	5	13	7	8	4	6		CGA	2	0	6	7	3	1	4	0
	CUG	1	5	15	22	6	3	5	4		CCG	2	0	8	12	1	1	4	4		CAG	5	1	3	9	9	3	6	8		CGG	3	1	5	8	3	7	2	3
Ile	AUU	5	2	14	18	8	2	4	8	Thr	ACU	2	4	4	6	1	2	3	6	Asn	AAU	9	4	8	13	4	4	7	8	Ser	AGU	3	1	0	8	1	0	1	2
	AUC	5	2	9	18	7	5	6	9		ACC	1	1	8	8	4	1	2	5		AAC	2	2	12	12	5	3	8	12		AGC	3	2	12	6	3	6	2	7
	AUA	7	2	11	9	1	1	2	5		ACA	5	3	9	14	3	2	2	2	Lys	AAA	8	4	12	17	4	6	0	6	Arg	AGA	1	1	7	5	1	1	3	3
Met	AUG	5	3	5	17	8	5	7	5		ACG	0	0	4	3	5	1	3	7		AAG	6	4	4	17	6	3	1	4		AGG	4	0	1	13	2	2	1	6
Val	GUU	9	1	9	14	3	4	3	8	Ala	GCU	9	1	13	19	6	7	3	10	Asp	GAU	7	2	16	23	6	9	9	6	Gly	GGU	1	2	9	19	4	7	4	6
	GUC	4	1	2	13	2	2	2	6		GCC	1	3	19	14	7	4	2	5		GAC	8	3	12	23	4	4	5	6		GGC	5	2	13	14	2	5	3	8
	GUA	1	2	11	3	0	1	3	3		GCA	3	3	13	18	6	1	5	14	Glu	GAA	7	4	13	23	6	7	9	10		GGA	2	2	13	16	9	3	7	6
	GUG	3	0	8	18	3	3	0	12		GCG	3	1	8	10	4	1	5	10		GAG	1	5	3	15	9	5	8	15		GGG	0	1	6	14	3	1	3	5

There is no general bias in the codon usage of these eight coding sequences taken together, although individually, large deviations do occur. We should note that the transcripts 1, 2, 3, 6a and 6b have a high preference for G as first base (>33.9%) and transcripts 4, 6a, 6b and 7 have a high percentage of A in the second position (>33.2%). No such deviations are noted in the third position.

be octopine synthase (transcript 3). The smallest one was selected with *Hind*III fragment 18 (Figure 1) and corresponds to the translated part of the gene transcript 7. The nucleotide sequences of both transcript 3 and 7 have been described (De Greve *et al.*, 1982a; Dhaese *et al.*, 1983). The third protein (mol. wt. = 27 kd) was observed after hybridization selection both with the partially overlapping fragments *Bam*HI-8 and *Hind*III-1 (Schröder and Schröder, 1982) (Figure 1). The authors suggested that at least part of the coding region is common to both fragments, but we do not find any open-reading frame in this part of the TL-region corresponding to a protein of this size. However, from Table III it appears that the polypeptides encoded by transcript 4 (located in *Hind*III fragment 1; Figure 1) and transcript 5 (located in *Bam*HI fragment 8; Figure 1) have nearly the same mol. wts. (26 873 and 25 635 daltons, respectively). The experimental results obtained by Schröder and Schröder (1982) can be explained if we assume that the observed 27-kd protein bands are in fact different and are encoded by transcripts 4 and 5, respectively.

The TL-region of octopine Ti plasmids expresses four proteins (mol. wt. = 74, 49, 28 and 27 kd) in *Escherichia coli* mini-cells (Schröder *et al.*, 1983). A comparison of the regions expressed in bacteria and the TL-region sequence indicates that three protein-coding regions in the bacteria correspond to three open-reading frames which are transcribed in plants (Table III). The mol. wts. of the polypeptides encoded by transcripts 2 (49 kd) and 4 (27 kd) as calculated from the sequence, are in good agreement with the mol. wts. experimentally observed by Schröder *et al.* (1983) in a bacterial background. However, there is a discrepancy between the calculated (84 kd) and the observed (74 kd) mol. wts. for the protein encoded by transcript 1. Schröder *et al.* (1983) showed that the right-end of the *Bam*HI-8 fragment (Figure 1) in pGV0153 encoded a 66-kd protein, which represents a shortened form of the 74-kd protein. The mol. wt. of this shortened protein calculated from the DNA sequence is 69 kd. Furthermore, deletion of fragment *Hpa*I-14, which is an internal fragment of *Eco*RI fragment 7 (Figure 1) that covers this region, produced a protein of mol. wt. = 53 kd (Schröder *et al.*, 1983). From the DNA sequence we can predict that the first 483 amino acids of transcript 1 will be fused to the last 16 amino acids of transcript 4 in this deletion mutant. The mol. wt. of this fusion protein is 55 kd, in good agreement with the mol. wt. (53 kd) observed by Schröder *et al.* (1983). It is likely, therefore, that the 74-kd protein is indeed encoded by the transcript 1 gene and that the difference in the observed and calculated mol. wts. can be explained by (i) an underestimation of the observed mol. wt. in SDS-polyacrylamide gels, or (ii) proteolytic degradation of this polypeptide in bacteria yielding a shorter protein.

Finally, Schröder *et al.* (1983) observed a 28-kd polypeptide in *E. coli* mini-cells. They located the gene encoding this polypeptide to the left of transcript 4. We do not find an open-reading frame in this region large enough to accommodate this 28-kd protein. Furthermore, no mRNA isolated from crown gall tumors has been observed to hybridize to this region.

Transcription initiation and polyadenylation signals. Comparisons of a multitude of eukaryotic protein-encoding genes have revealed a limited number of consensus sequences potentially involved in RNA polymerase II-mediated transcription. The 'TATA' box or Goldberg-Hogness box (Proudfoot, 1979) is located 25–30 bp upstream from the start site of transcription and is involved *in vivo* in the accurate positioning of the mRNA start site (McKnight and Kingsbury, 1982). The consensus sequence GG(C/T)CAATCT of 'CCAAT' box (Benoist *et al.*, 1980), which appears 40–50 nucleotides upstream of the TATA box, is involved in the regulation of transcription of some eukaryotic genes. By comparing plant genes, a possible regulatory sequence, called AGGA box, was identified by Messing *et al.* (1983). As the transcription of TL-DNA genes is α-amanitin sensitive (Willmitzer *et al.*, 1981) and potential control signals in the 5′ regions of the T-DNA genes (De Greve *et al.*, 1982a; Depicker *et al.*, 1982; Dhaese *et al.*, 1983; Heidekamp *et al.*, 1983), of which the transcription initiation site was accurately determined, have been found resembling those typically used by eukaryotes, we

Table V. Eukaryotic signals present in 5′ and 3′ sequences of the different transcripts

	Position	'CCAAT' box	Position	'TATA' box	Position	Poly(A)$^+$
Consensus sequence		GG$^{C}_{T}$CAATCT		TATA$^{A}_{T}$A$^{A}_{T}$		AATAAA
Transcript 5	909	GGCgAATaT	983	aATAAtA	1912	AATAAT
	935	acgCAATta	1012	TATAAgA	1948	AATAAT
	979	taCCAATaa	1029	TtTATAT		
	1001	GGCCAtTta				
Transcript 7	2800	GtTCAAgCT	2735	TATATAT	2188	AATAAA
Transcript 2	4932	GcgCAAgCT	4909	TATATtT	3281	AATAAT
	4943	caCCAATaa			3297	AATAAT
					3312	AATAAA
					3364	AATAAT
Transcript 1	5092	GcCCAAatT	5175	TATtTAT	7710	AATAAT
	5118	tGTCAAcga			7727	AATAAT
	5144	tcTCAActT				
Transcript 4	8072	ctTCAATaa	8098	aATATAA	9101	AATAAA
	8080	aaTgAATtT	8131	TATAAAA	9169	AATAAA
	8094	aGaCAATaT				
Transcript 6a	9294	GcgaAATtT	9326	TATtAAT	10 030	TATAAA
					10 085	AATGAA
Transcript 6b	11 169	caCCAATga	11 137	TATAAAA	10 260	AATAAT
	11 204	taTCAATCT			10 355	AATAAA
					10 434	AATAAA
Transcript 3	13 114	aCTCAATac	13 088	TATtTAA	11 778	AATAAT
					11 810	AATATA
					11 814	AATGAA

searched for homologies with these putative regulatory sequences in the 5′-untranslated region of the TL-DNA genes. In the 5′-untranslated region of transcript 5, three sequences AATAATA, TATAAGA, and TTTATAT (position 983, 1012 and 1029), sharing homology with the TATA sequence, are located respectively 77, 48 and 31 bp upstream from the translation start codon and are preceded by four 'CCAAT'-like sequences (GGCGAATAT at position 909, ACGCAATTA at 935, TACCAATAA at 979, GGCCATTTA at 1001). Transcript 2 has a TATATTT sequence (position 3460) and two possible CCAAT sequences (GCGCAAGCT at position 4932 and CACCAATAA at 4943). A TATTTAT sequence (position 5175) is located 34 bp upstream from the translation start codon of the gene encoding transcript 1. This TATA box is preceded by three possible CCAAT boxes (positions 5692, 5118, and 5114). The 5′-untranslated region of the gene encoding transcript 6a contains a TATTAAT sequence (position 9326) located 69 bp upstream from the ATG translation codon and a CCAAT sequence (position 9294) located 32 bp upstream from the presumed TATA box. The gene encoding transcript 6b has a TATAAAA sequence (position 11 137) 61 bp upstream from the translation start codon. Two CCAAT sequences (position 11 169 and 11 204) are located upstream of the TATA box at a distance of 32 bp and 67 bp. A summary of the eukaryotic signals found in the 5′-untranslated regions is listed in Table V. However, we did not find sequences in the 5′-untranslated regions of the TL-DNA sharing significant homology with the AGGA box (Messing *et al.*, 1983).

Sequences essential for the *in vivo* expression of eukaryotic genes, however, are located, in most cases, 200–300 bp upstream of the transcription initiation site. From genetic studies, there is evidence that sequences upstream of the TATA and CCAAT boxes are also involved in the *in vivo* expression of the octopine synthase gene (Koncz *et al.*, 1983) in plant cells. We did not find nucleotide sequence homology between this 5′ upstream region of the octopine synthase gene and the 5′ upstream regions of the other TL-DNA genes.

Most eukaryotic protein-encoding transcripts are polyadenylated. The only primary sequence common to the 3′-untranslated region of almost all eukaryotic genes is the hexanucleotide AATAAA (Proudfoot and Brownlee, 1976; Benoist *et al.*, 1980), or a one-base variation of this sequence (Nevins, 1983). This sequence functions in the recognition of the poly(A) addition site (Fitzgerald and Shenk, 1981; Montell *et al.*, 1983). The poly(A) addition sites of the octopine synthase (De Greve *et al.*, 1982a), the nopaline synthase (Depicker *et al.*, 1982), the octopine synthase present in the regenerated plant rGV1 and transcript 7 (Dhaese *et al.*, 1983) are indeed closely preceded by this hexanucleotide signal. In the case of the wild-type octopine synthase and the rGV1 octopine synthase multiple polyadenylation sites have been observed. This was also found to occur in animal genes

(Setzer *et al.*, 1980; Early *et al.*, 1980). We looked for the presence of AATAAA or related sequences in the 3'-untranslated regions of the TL-DNA genes encoding transcripts 5, 2, 1, 6a and 6b. For each gene at least two potential canonical sequences are found. Transcripts 5 and 1 each contain two polyadenylation signals AATAAT (position 1912 and 1948 for transcript 5 and 7710 and 7727 for transcript 1). In transcript 5, these are located at a distance of 172 bp and 208 bp downstream of the stop codon, and those of transcript 1 at 234 bp and 251 bp downstream from the stop codon. The 3'-untranslated region of transcript 2 contains four possible polyadenylation signals: AATAAT (position 3281), AATAAT (3297), AATAAA (3312) and AATAAT (3364), respectively 96, 148, 163 and 180 bp, past the translational stop. In the 3' region of transcript 6b three polyadenylation signals AATAAT (10 260), AATAAA (10 355), and AATAAA (10 434) are found respectively 193, 98 and 19 bp downstream from the stop codon. Transcript 6a has two sequences: TATAAA (10 030) and AATGAA (10 085) in its 3' end which are located at a distance of 60 bp and 115 bp downstream from the stop codon. All these data are summarized in Table V.

Translation initiation codons. In eukaryotes, the first AUG of the majority of mRNAs is used as an initiation codon. In the scanning model, two bases (A or G at position −3, G at position +4) flanking the initiation codon (A/GXXAUGG) facilitate the recognition of the functional AUG codon (Kozak, 1981).

Since none of the amino acid sequences of the proteins encoded by the TL-DNA in plant cells have been determined, no experimental data exist concerning the sites used to initiate translation of the plant transcripts. As can be seen in Figure 2, the first AUG following the 'TATA' box is in phase with all the open-reading frames and most likely initiates translation in plants. The first AUG of these plant transcripts are preceded by a very G-poor stretch of DNA and do not contain a Shine-Dalgarno sequence (Shine and Dalgarno, 1974; Stormo *et al.*, 1982). This lack of Gs upstream of eukaryotic initiation codons has already been observed (Kozak, 1981; Sargan *et al.*, 1982).

In the open-reading frames of the genes encoding transcript 5, 7, 2, 4 and 3 the second AUG is located at a distance of 300, 231, 354 and 252 bp, respectively, of the first AUG. In the case of open-reading frames 2 and 4, which are translated in *E. coli* mini-cells (Schröder *et al.*, 1983) these data support the hypothesis that the same translational start is used in bacteria as well as in plant cells. Two AUG codons (positions 11 019 and 11 076) can be used as initiation codon for transcript 6b. Both AUG codons are flanked by a G (position −3) and an A (position +4). Because the initiation codons are equivalent, there is no reason to believe that the first AUG codon is not used as the translational start.

In transcript 6a three AUG codons (position 9395, 9404 and 9410) can be used as initiation codon. The first and the third AUG codons are flanked by two bases which facilitate the recognition of functional AUG codons (Kozak, 1981). Comparison of the TL-DNA sequence of transcript 6a with the corresponding nopaline T-DNA sequence (unpublished data) indicate that in the homologous pTiC58 sequence only the third AUG is conserved. This observation suggests that translation of the octopine transcript 6a starts at the third AUG. However, we cannot exclude that the transcripts 6a encoded by the octopine TL-DNA and the nopaline T-DNA, respectively, have different translational starts.

Transcript 1 also contains three AUG condons in the beginning of the frame (positions 5209, 5260 and 5275). Although we have no data to support that the first AUG is not used as the initiation signal in the plant cells, the possibility exists that the third AUG, which is preceded by a GGTGGA sequence (position 5262) might be preferably used in a bacterial background. The difference in mol. wt. will be 2.3 kd, when calculated from the sequence, and the correspondence with the observed mol. wts. of the shorter polypeptides (53 and 66 kd) (Schröder *et al.*, 1983) and the computed mol. wts. (52.7 and 66.7 kd) are even better.

To solve the question of whether the same translation start codon is used in plant cells and in bacteria, amino acid sequences of both will be needed.

Intervening sequences. A characteristic but not an absolute criterion of eukaryotic genes is the presence of intervening se-

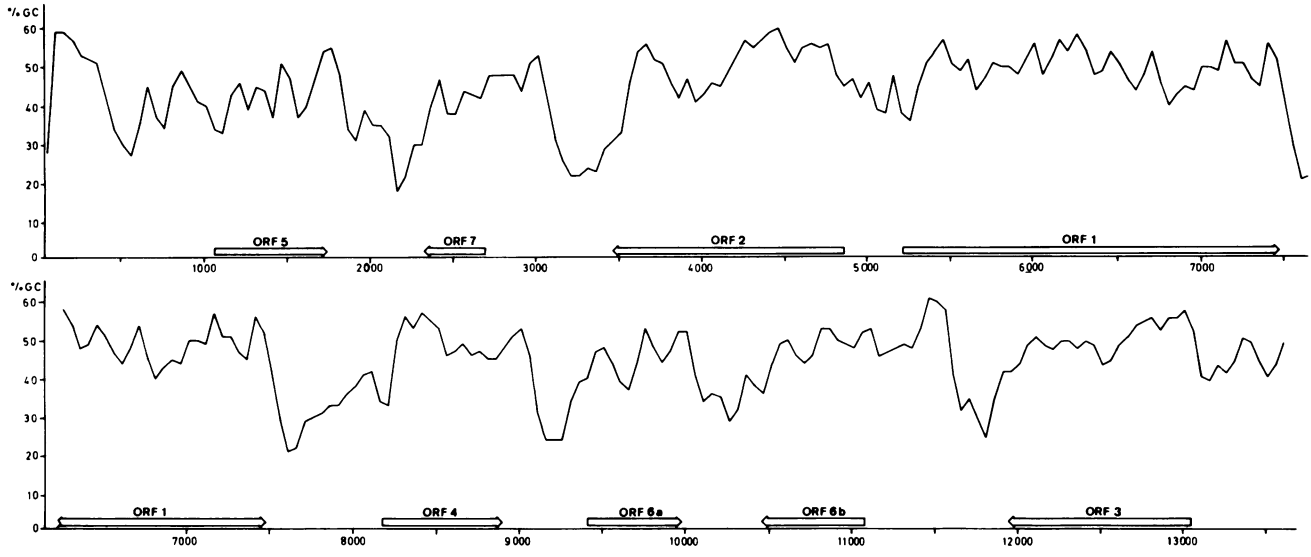

Fig. 4. GC profile of the TL-DNA. A window of 100 bp was slid along the sequence by increments of 50 bp, and its G+C percentage was calculated. The position and size of each known coding region and its orientation is indicated by arrows. The two parts of the figure are contiguous, but the right part of transcript 1 is repeated in the lower figure in order to emphasize the periodicity of the GC content.

quences. To date, several plant nuclear genes have been shown to contain intervening sequences (Sun *et al.*, 1981; Fisher and Goldberg, 1982; Hyldig-Nielsen *et al.*, 1982; Shah *et al.*, 1982), while several others lack intervening sequences (Geraghty *et al.*, 1981; Fisher and Goldberg, 1982; Pedersen *et al.*, 1982). The existence of introns in the coding regions of the different TL-DNA transcripts is very unlikely. Firstly, the open-reading frames correlate well with the sizes of the cytoplasmic polyadenylated transcripts 1, 2, 3, 4, 5, 6a, 6b and 7, determined by Northern analysis (Willmitzer *et al.*, 1982, 1983). Secondly, as discussed above, the sizes of the proteins observed experimentally *in vitro* (Schröder and Schröder, 1982), and in *E. coli* (Schröder *et al.*, 1983) correspond nicely to those calculated from the sequence presented in Figure 3. Furthermore, we have looked without success for sequences fitting with the donor and acceptor consensus sequences proposed by Mount (1982) normally found at the intron-exon junctions.

G+C content. A striking feature of the TL-DNA sequence (Figure 4) is observed when a graphical display of a G+C content profile is plotted. Each functional coding sequence is separated from its neighbours by an AT-rich interval. The 3'-untranslated region of each transcript is very AT-rich, a feature also observed in the 3'-untranslated region of other plant genes, ranging from 24% G+C in the soybean leghemoglobin gene (Hyldig-Nielsen *et al.*, 1982) to 37% G+C in the ribulose-1,5-biphosphate carboxylase gene (Bedbrook *et al.*, 1980). The dip in the G+C profile is less marked between transcripts 1 and 2, possibly because in this case both 5' ends are very close to one another. Furthermore, these large variations of G+C content can be visualized under the electron microscope by partial denaturation of the Ti plasmid and are limited to the TL-region and the homologous region of the nopaline T-DNA (G. Engler, personal communication).

Conclusions

From the determination and the analysis of the primary structure of the TL-DNA sequence, the following conclusions can be drawn: (i) all the TL-DNA genes contain the signals to be transcribed and translated in plant cells; (ii) the absence of intervening sequences and the compact organization of the genes on the TL-DNA suggest that a maximum amount of genetic information is concentrated in a minimum amount of DNA.

Materials and methods

Enzymes

DNA polymerase I (large fragment, according to Klenow) and T4 polynucleotide kinase were from Boehringer Pharma (Mannheim, FRG).

Restriction enzymes were from Boehringer Pharma (Mannheim, FRG) or New England Biolabs (Beverly, MA, USA), and were used according to the suppliers' instructions.

Bacterial strains and plasmids

Bacterial strains and plasmids are listed in Table I.

Plasmid preparation

Agarose gel electrophoresis, conditions for DNA ligation, and transformation of competent *E. coli* cells were as described by Depicker *et al.* (1980).

Plasmids were prepared from *E. coli* K514 or by CsCl-EtBr equilibrium density gradient centrifugation in cleared SDS lysates (Betlach *et al.*, 1976). The copy number of the pBR derivatives was increased by adding chloramphenicol (170 μg/ml) or spectinomycin (300 μg/ml) to an exponentially growing culture and incubating for a further 15 h.

DNA sequence determination

DNA fragments to be sequenced were labeled at their 5' ends with [γ-^{32}P]-ATP (>2000 Ci/mmol, Amersham) and T4 polynucleotide kinase (Boehringer, Mannheim, FRG) after treatment with bacterial alkaline phosphatase (Boehringer, Mannheim, FRG); DNA fragments wre labeled at their 3' ends using either [^{32}P]cordycepin (NEN) and terminal nucleotidyl transferase, or [α-^{32}P]dATP and Klenow polymerase (Boehringer, Mannheim, FRG). The labeled fragments, after secondary restriction, were extracted from low-gelling temperature agarose as described by Wieslander (1979), or, after strand separation, were extracted from acrylamide as described by Maxam and Gilbert (1980).

The five chemical modification and cleavage reactions G, A+G, C+T, C and A+C were performed as described by Maxam and Gilbert (1980). The cleavage products were separated on 8% and 15% gradient acrylamide gels (0.3 mm x 90 cm) containing 8.3 M urea (Sanger and Coulson, 1978). The gels were autoradiographed at −70°C using intensifying screens.

Computer analysis

Routine analysis (restriction sites, overlaps) of the sequencing data was performed on a Cromemco microcomputer using the mapping and comparison programs written by Schroeder and Blattner (1982) for the CP/M operating system. We developed a program along the lines of the RNY algorithm, described by Shepherd (1981) and the programs used to calculate the mol. wt. of the proteins (Table II), the codon usage (Table III), and the GC profile of the sequence (Figure 4). The limited computing ability of our microcomputer did not allow us to perform extensive searches of similarities using the Sellers (1979), or Needleman and Wunsch (1970) algorithms. Imperfect repeats might therefore have escaped. A machine-readable copy of the sequence has been sent for incorporation in the EMBL data base.

Acknowledgements

The authors wish to thank Ms M. De Cock for typing the complete nucleotide sequence and the manuscript, and K. Spruyt and A. Verstraete for photographical work. This research was supported by grants of the 'ASLK-Kankerfonds', 'Fonds voor Geneeskundig Wetenschappelijk Onderzoek' (FGWO 3.0001.82), 'Instituut tot aanmoediging van het Wetenschappelijk Onderzoek in Nijverheid en Landbouw' (IWONL 3849A), and the Services of the Prime Minister (OOA 12052179), and was carried out under Research Contract No. GBI-4-017-B(RS) of the Biomolecular Engineering Programme of the Commission of the European Communities.

References

Bedbrook,J.R., Smith,S.M. and Ellis,R.J. (1980) *Nature,* **287**, 692-697.
Benoist,C., O'Hare,K., Breathnach,R. and Chambon,P. (1980) *Nucleic Acids Res.,* **8**, 127-142.
Betlach,M.C., Hershfield,V., Chow,L., Brown,W., Goodman,H.M. and Boyer,H.W. (1976) *Fed. Am. Soc. Exp. Biol.,* **35**, 2037-2043.
Bevan,M.W. and Chilton,M.-D. (1982) *J. Mol. Appl. Genet.,* **1**, 539-546.
Caplan,A., Herrera-Estrella,L., Inzé,D., Van Haute,E., Van Montagu,M., Schell,J. and Zambryski,P. (1983) *Science (Wash.),* **222**, 815-821.
Chilton,M.-D., Drummond,M.H., Merlo,D.J., Sciaky,D., Montoya,A.L., Gordon,M.P. and Nester,E.W. (1977) *Cell,* **11**, 263-271.
Chilton,M.-D., Drummond,M.H., Merlo,D.J. and Sciaky,D. (1978) *Nature,* **275**, 147-149.
Colson,C., Glover,S.W., Symonds,N. and Stacey,K.A. (1965) *Genetics,* **52**, 1043-1050.
De Beuckeleer,M., Lemmers,M., De Vos,G., Willmitzer,L., Van Montagu, M. and Schell,J. (1981) *Mol. Gen. Genet.,* **183**, 283-288.
De Greve,H., Dhaese,P., Seurinck,J., Lemmers,M., Van Montagu,M. and Schell,J. (1982a) *J. Mol. Appl. Genet.,* **1**, 499-512.
De Greve,H., Leemans,J., Hernalsteens,J.P., Thia-Toong,L., De Beuckeleer,M., Willmitzer,L., Otten,L., Van Montagu,M. and Schell,J. (1982b) *Nature,* **300**, 752-755.
Depicker,A., Van Montagu,M. and Schell,J. (1978) *Nature,* **275**, 150-153.
Depicker,A., De Wilde,M., De Vos,G., De Vos,R., Van Montagu,M. and Schell,J. (1980) *Plasmid,* **3**, 193-211.
Depicker,A., Stachel,S., Dhaese,P., Zambryski,P. and Goodman,H.M. (1982) *J. Mol. Appl. Genet.,* **1**, 561-574.
De Vos,G., De Beuckeleer,M., Van Montagu,M. and Schell,J. (1981) *Plasmid,* **6**, 249-253.
Dhaese,P., De Greve,H., Gielen,J., Seurinck,J., Van Montagu,M. and Schell,J. (1983) *EMBO J.,* **2**, 419-426.
Early,P., Rogers,J., Davis,M., Calame,K., Bond,M., Wall,R. and Hood,L. (1980) *Cell,* **20**, 313-319.
Engler,G., Depicker,A., Maenhaut,R., Villarroel-Mandiola,R., Van Montagu,M. and Schell,J. (1981) *J. Mol. Biol.,* **152**, 183-208.
Fisher,R.L. and Goldberg,R.B. (1982) *Cell,* **29**, 651-660.
Fitzgerald,M. and Shenk,T. (1981) *Cell,* **24**, 251-260.
Garfinkel,D.J., Simpson,R.B., Ream,L.W., White,F.F., Gordon,M.P. and

Nester,E.W. (1981) *Cell,* **27**, 143-153.
Geraghty,D., Peifer,M.A., Rubenstein,I. and Messing,J. (1981) *Nucleic Acids Res.,* **9**, 5163-5174.
Heidekamp,F., Dirkse,W.G., Hille,J. and von Ormondt,H. (1983) *Nucleic Acids Res.,* **11**, 6211-6223.
Holsters,M., Villarroel,R., Gielen,J., Seurinck,J., De Greve,H., Van Montagu,M. and Schell,J. (1983) *Mol. Gen. Genet.,* **190**, 35-41.
Hyldig-Nielsen,J.J., Jensen,E.Ø., Paludan,K., Wiborg,O., Garrett,R., Jørgensen,P. and Marcker,K.A. (1982) *Nucleic Acids Res.,* **10**, 689-695.
Joos,H., Inzé,D., Caplan,A., Sormann,M., Van Montagu,M. and Schell,J. (1983) *Cell,* **32**, 1057-1067.
Kahl,G. and Schell,J. (1982) *Molecular Biology of Plant Tumors*, published by Academic Press, NY, 615 pp.
Koncz,C., De Greve,H., André,D., Deboeck,F., Van Montagu,M. and Schell,J. (1983) *EMBO J.,* **2**, 1597-1603.
Kozak,M. (1981) *Nucleic Acids Res.,* **9**, 5233-5252.
Leemans,J., Deblaere,R., Willmitzer,L., De Greve,H., Hernalsteens,J.P., Van Montagu,M. and Schell,J. (1982) *EMBO J.,* **1**, 147-152.
Lemmers,M., De Beuckeleer,M., Holsters,M., Zambryski,P., Depicker,A., Hernalsteens,J.P., Van Montagu,M. and Schell,J. (1980) *J. Mol. Biol.,* **144**, 353-376.
Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.,* **65**, 499-559.
McKnight,S.L. and Kingsbury,R. (1982) *Science (Wash.),* **217**, 316-324.
Messing,J., Geraghty,D., Heidecker,G., Hu,N.-T., Kridl,J. and Rubenstein, I. (1983) in Hollaender,A. (ed.), *Genetic Engineering of Plants*, Plenum Press, NY, pp. 211-227.
Montell,C., Fisher,E.F., Caruthers,M.H. and Berle,A.J. (1983) *Nature,* **305**, 600-605.
Mount,S.M. (1982) *Nucleic Acids Res.,* **10**, 459-472.
Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.,* **48**, 443-453.
Nester,E.W. and Kosuge,T. (1981) *Annu. Rev. Microbiol.,* **35**, 531-565.
Nevins,J.R. (1983) *Biochemistry (Wash.),* **52**, 441-466.
Ohmori,H., Tomizawa,J. and Maxam,A.M. (1978) *Nucleic Acids Res.,* **5**, 1479-1485.
Ooms,G., Bakker,A., Molendijk,L., Wullems,G.J., Gordon,M.P., Nester, E.W. and Schilperoort,R.A. (1982) *Cell,* **30**, 589-597.
Pedersen,K., Devereux,J., Wilson,D.R., Sheldon,E. and Larkins,B.A. (1982) *Cell,* **29**, 1015-1026.
Proudfoot,N.J. (1979) *Nature,* **279**, 376.
Proudfoot,N.J. and Brownlee,G.G. (1976) *Nature,* **263**, 211-214.
Sanger,F. and Coulson,A.R. (1978) *FEBS Lett.,* **87**, 107-110.
Sargan,D.R., Gregory,S.P. and Butterworth,P.H.W. (1982) *FEBS Lett.,* **147**, 133-136.
Schroeder,J.L. and Blattner,F.R. (1982) *Nucleic Acids Res.,* **10**, 69-84.
Schröder,G. and Schröder,J. (1982) *Mol. Gen. Genet.,* **185**, 51-55.
Schröder,J., Hillebrand,A., Klipp,W. and Pühler,A. (1981) *Nucleic Acids Res.,* **9**, 5187-5202.
Schröder,G., Klipp,W., Hillebrand,A., Ehring,R., Koncz,C. and Schröder,J. (1983) *EMBO J.,* **2**, 403-409.
Sellers,P.H. (1979) *Proc. Natl. Acad. Sci. USA,* **76**, 3041.
Setzer,D.R., McGrogan,M., Nunberg,J.H. and Schimke,R.T. (1980) *Cell,* **22**, 361-370.
Shah,D.M., Hightower,R.C. and Meagher,R.D. (1982) *Proc. Natl. Acad. Sci. USA,* **79**, 1022-1026.
Shepherd,J.C. (1981) *Proc. Natl. Acad. Sci. USA,* **78**, 1596-1600.
Shine,J. and Dalgarno,L. (1974) *Proc. Natl. Acad. Sci. USA,* **71**, 1342-1346.
Simpson,R.B., O'Hara,P.J., Krook,W., Montoya,A.L., Lichtenstein,C., Gordon,M.P. and Nester,E.W. (1982) *Cell,* **29**, 1005-1014.
Stormo,G.D., Schneider,T.D. and Gold,L.M. (1982) *Nucleic Acids Res.,* **10**, 2971-2996.
Sun,S.M., Slightom,J.L. and Hall,T.C. (1981) *Nature,* **289**, 37-41.
Thomashow,M.F., Nutter,R., Montoya,A.L., Gordon,M.P. and Nester, E.W. (1980) *Cell,* **19**, 729-739.
Wieslander,L. (1979) *Anal. Biochem.,* **98**, 305-309.
Willmitzer,L., Schmalenbach,W. and Schell,J. (1981) *Nucleic Acids Res.,* **9**, 4801-4812.
Willmitzer,L., Simons,G. and Schell,J. (1982) *EMBO J.,* **1**, 139-146.
Willmitzer,L., Dhaese,P., Schreier,P.H., Schmalenbach,W., Van Montagu, M. and Schell,J. (1983) *Cell,* **32**, 1045-1056.
Yadav,N.S., Vanderleyden,J., Bennett,D.R., Barnes,W.M. and Chilton, M.-D. (1982) *Proc. Natl. Acad. Sci. USA,* **79**, 6322-6326.
Zambryski,P., Depicker,A., Kruger,K. and Goodman,H. (1982) *J. Mol. Appl. Genet.,* **1**, 361-370.
Zambryski,P., Goodman,H., Van Montagu,M. and Schell,J. (1983) in Shapiro,J.A. (ed.), *Mobile Genetic Elements,* Academic Press, NY, pp. 505-535.

Received on 2 January 1984