

**PROMOTER TRAPPING IN *ARABIDOPSIS THALIANA*:
CHARACTERIZATION OF T-DNA TAGGED LINES**

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Pedro Salvador Carteado Ferreira da Rocha

(Lic., Porto; M.Sc, Guelph)

Department of Biology

University of Leicester

March 1999

UMI Number: U535929

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U535929

Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Promoter trapping in *Arabidopsis thaliana*: Characterization of T-DNA tagged lines

by

Pedro Salvador Carteadó Ferreira da Rocha

Two *A. thaliana* lines transgenic for the promoter-trap vector p Δ *gus*Bin19 were studied. Line AtEN101 lacked a phenotype but reporter gene (*gus*) expression exhibited temporal and spatial regulation, including in the developing embryo. This pattern was unaltered when AtEN101 was used as a functional marker in two embryo-defective mutant backgrounds tested. In roots of AtEN101 seedlings, *gus* expression was repressed by ABA and modified by auxins, in a sucrose-modulated manner.

The tagged genomic region in the mutant and wildtype, and cDNAs mapping to it were isolated. The cDNAs derived from two overlapping genes (*PKT1* and *PKT2*) almost identical except for their first exons and introns and encoding putative peroxisomal 3-ketoacyl-CoA thiolases. The T-DNA was inserted in the long intron1 (966 bp) of *PKT1*, and 0.3 kb upstream from exon1 of *PKT2*. *gus* copies were found at both T-DNA junctions. Fusion transcripts including exon1 of *PKT1* and T-DNA sequences were identified, and a cryptic 3' splice site identified in the left border region.

The genes were found to be independently expressed, at low levels, in all organs by Northern blot hybridization and RT-PCR. Both putative promoter regions were capable of driving the expression of *gus* in transient assays. That of *PKT1* was also functional in reverse orientation. The thiolase genes are part of a family of at least four members, *PKT1* being unique in lacking the characteristic N-terminal peroxisomal targeting signal, PTS2.

A new T-DNA linked semi-dominant mutation causing dwarfism (*bashful*) was identified in line P26K4. *gus*-derived sequences were present in its genome but no GUS activity was observed. The mutation affected cell growth and was rescued by epibrassinolide but not gibberellic acid (GA₃). Skotomorphogenic development is also affected in the mutant.

The utility of p Δ *gus*Bin19 as a promoter-trap and insertional mutagen in *Arabidopsis*, and the possible functions of *PKT1*/*PKT2* are discussed.

Para a Mamã

Acknowledgments

I would like to thank my supervisor Professor Keith Lindsey for the opportunity to conduct the research and all his support and guidance throughout.

My thanks are also extended to Dr. D. Twell for accepting to be my administrative supervisor after my moving to Durham.

I am grateful to past and present colleagues for all their support, in particular Dr. J.F. Topping, who supplied many of the materials used, not least of which AtEN101. Also special thanks are due to Dr. W. Wei particularly for the always useful discussions. I am also indebted to Mr. S. Casson, Ms. K. Farrar, and Dr. M. Evans for sharing some of their unpublished results. A thank you also goes to Mr. Graham at the Leicester Botanical Gardens for his good care of some of the plants used in this work.

I am also grateful to all those who supplied materials used in this work including Dr. M.-K. Cheung for plasmid pMKC1, Professor N.-H. Chua for the seeds of *diminuto*, Drs. Dangel and Davies for the λ genomic library, Dr. J. Gatehouse for the tobacco SR1 seeds, Dr. J. Giraudat for the λ cDNA library, Dr. M. Hooks for the Northern blot of germinating seedlings RNAs, Professor M. Hughes for access to the particle gun, Dr. Sedgwick for the Tn Δ XR transposon and relevant *E. coli* strains, and Dr. A. Sorensen for the *A. thaliana* cv. Columbia seeds.

Thanks are also due to Dr. W. Cockburn and Dr. I. Graham for agreeing to examine this thesis.

Last but not least I am very grateful to the Junta Nacional de Investigação Científica e Tecnológica for the financial support without which this work could not have been realized.

Table of Contents

Table of contents	i
Abbreviations	xiii
Chapter 1: Introduction	1
1.1 An historical perspective on the study of development	1
1.1.1 Beginnings	1
1.1.2 Development and Heredity	3
1.2 Embryogenesis in higher plants	5
1.2.1 Generalities	5
1.2.2 <i>Arabidopsis</i> as a model biological system	7
1.2.3 Zygotic embryogenesis in <i>A. thaliana</i>	9
1.2.4 Gene activity and embryogenesis	10
1.2.4.1 Genic autonomy of the developing embryo	10
1.2.4.2 Approaches to identify embryonically active genes	11
1.2.4.3 Regulatory complexities of the embryo gene expression program	13
1.2.4.4 Regulatory mechanisms of gene expression	14
1.3 Current methodology for identification and cloning of genes	16
1.3.1 Strategies based on the products of gene expression	17
1.3.1.1 RNA-based methods	17
1.3.1.1.1 Direct isolation	17
1.3.1.1.2 Use of differential expression	18
a) Differential screening	
b) Cloning by subtraction hybridization	
c) Differential display	
1.3.1.2 Cloning genes from their protein products	20
1.3.2 Homology-based cloning	22
1.3.2.1 Screening with homologous genes and cDNAs	22
1.3.2.2 Two uses for oligonucleotides	23
1.3.2.3 Antibody screening	23

1.3.2.4	Homology-based selection of ESTs	24
1.3.3	Genetic/Functional strategies	24
1.3.3.1	Complementation cloning	24
1.3.3.2	Dihybrid system for detection of interacting proteins	24
1.3.3.3	Monohybrid system	25
1.3.3.4	Phage display and related techniques	25
1.3.4	Methods based on chromosomal differences	26
1.3.4.1	Cloning of deleted genes	26
	a) Positional cloning	
	b) Molecular markers linkage and high density mapping	
	c) Chromosomal landing and synteny-based mapping	
	d) Applications in plants	
1.3.4.2	Representational difference analysis (RDA)	26
1.3.4.3	Cloning genes from their physical location	27
1.3.5	Gene tagging by insertional mutagenesis	29
1.3.5.1	Introduction	29
1.3.5.2	Transposons as molecular tags in plants	30
1.3.5.3	T-DNA tagging	31
1.3.5.4	Promoter- and enhancer-trapping	33
1.3.5.5	Activation tagging	36
1.4	Objectives of the research	36
1.4.1	Promoter-trapping and embryogenesis	36
1.4.2	Specific objectives of the research	38

Chapter 2: Materials and Methods 39

2.1	Materials	39
2.1.1	Chemicals and Reagents	39
2.1.1.1	General sources	39
2.1.1.2	Water	39
2.1.1.3	Preparation of phenol solutions	39
2.1.2	Biochemicals	39
2.1.2.1	Radioisotopes	39
2.1.2.2	Oligodeoxynucleotides	40

2.1.2.3	Enzymes, other proteins and molecular biology kits	40
2.1.2.4	Antibiotics and other biochemicals	40
2.1.2.5	Plasmids	41
2.1.2.6	Libraries of clones	42
2.1.3	Bacterial strains and media	42
2.1.3.1	Bacterial strains	42
2.1.3.2	Bacterial media	43
2.1.4	Plant material and growth	44
2.1.4.1	Plant material	44
2.1.4.2	Plant culture media	45
2.1.4.3	Hormone and antibiotic preparation and use	45
2.1.4.4	Seed sterilization, germination and growth	45
2.2	Bacterial methods	46
2.2.1	Growth and storage of bacterial strains	46
2.2.2	Transformation of bacteria	47
2.2.3	Generation of transposon-mutagenized plasmids	48
2.3	Bacteriophage methods	48
2.3.1	Plating of λ phage	48
2.3.1.1	Titering	49
2.3.1.2	Preparation of λ stocks	49
2.3.2	Screening of libraries in λ phage	49
2.3.3	Excision of phagemid from λ ZAPII	51
2.4	DNA isolation	51
2.4.1	Plasmid DNA isolation	51
2.4.1.1	Mini-preparations	51
2.4.1.2	Large-scale preparations	53
2.4.2	λ phage DNA	54
2.4.2.1	Large-scale isolation	54
2.4.2.2	Small-scale λ DNA isolation	55
2.4.3	Plant DNA isolation	56
	CTAB method	56
2.4.4	Purification of DNA with Wizard TM clean-up system	58
2.4.5	Purification of oligodeoxynucleotides	58
2.4.6	Purification of DNA from gels	58

2.4.6.1	Electroelution	58
2.4.6.2	Purification with GeneClean II®	59
2.4.7	Purification of PCR products	59
2.5	Manipulation and analysis of DNA	60
2.5.1	Enzymatic manipulation	60
2.5.1.1	Restriction endonuclease digestions	60
2.5.1.2	Filling-in of recessive ends	60
2.5.1.3	Dephosphorylation of 5' ends	60
2.5.1.4	Ligation of DNA fragments	61
2.5.2	DNA agarose gel electrophoresis	61
2.5.3	Southern blot preparation	62
2.5.4	Sequencing of DNA	62
2.5.4.1	Manual sequencing	62
2.5.4.2	Automated sequencing	64
2.5.5	PCR amplification	65
2.5.5.1	General procedures	65
2.5.5.2	Amplification of genomic DNA	65
2.5.5.3	Cloning of PCR products	65
2.5.5.4	Colony-PCR	66
2.5.6	Assembly of <i>c-myc</i> fragment	66
2.6	RNA isolation and purification	67
2.6.1	Reducing RNAase contamination	67
2.6.2	Phenol/SDS method	67
2.6.3	PIC method	68
2.6.4	Isolation of poly(A) ⁺ RNA	70
2.7	RNA manipulation and analysis	70
2.7.1	Denaturing agarose gel electrophoresis	70
2.7.2	Northern blot preparation	71
2.7.3	RT-PCR	71
2.7.4	RACE of 5' ends of transcripts	72
2.7.4.1	SLIC method	72
2.7.4.2	LOPRACE	76
2.7.4.3	5' RACE with homopolymeric tailing	77
2.8	Nucleic acid hybridization	79

2.8.1	Labeling of probes	79
2.8.2	Probe purification	80
2.8.3	Southern blot hybridization	80
2.8.4	Northern blot hybridization	81
2.8.5	Autoradiography and phosphoimaging	82
2.9	Biolistics for functional assays of promoter activity	82
2.10	Detection of GUS enzyme activity	84
2.10.1	Fluorimetric detection	84
2.10.2	Histochemical detection of GUS activity	85
2.11	Computer-based searches, analysis and modelling	86

Chapter 3: Screening and phenotypic characterization of mutants 88

3.1	Introduction	88
3.2	Line AtEN101	88
3.2.1	Previous studies	88
3.2.2	Fluorimetric studies	89
3.2.3	Histochemical characterization	89
3.2.4	Phenotypical characterization	91
3.3	Identification and characterization of the <i>bashful</i> mutant	91
3.3.1	Screening for mutations affecting seedling pattern and shape	91
3.3.2	The <i>bashful</i> phenotype	92
3.3.3	Skotomorphogenic development in <i>bashful</i>	93
3.3.4	<i>bashful</i> is not allelic to <i>diminuto</i>	94
3.3.5	The <i>bashful</i> phenotype can be rescued by brassinosteroids	95
3.3.6	Linkage of the <i>bashful</i> phenotype to a T-DNA locus	95
3.3.7	Presence of <i>gus</i> reporter genes in <i>bashful</i>	96
3.4	Summary	100

Chapter 4: Mapping transcripts to the T-DNA insertion region

4.1	Introduction	102
4.2	Characterization of the T-DNA tagged region of AtEN101	103
4.2.1	Selection of an IPCR clone	103
4.2.2	Sequence at the T-DNA border	103
4.2.3	Building a restriction and features map for pIPCR2	103
4.2.4	Northern blotting hybridization	105
4.2.5	Sequencing of the pIPCR2 insert	105
4.2.6	Search for protein-coding regions	107
4.2.6.1	GeneBank searches	107
4.2.6.2	Splice sites at the junction region	108
4.2.6.3	Potential splice sites in the T-DNA left border region	108
4.3	The wildtype genome at the T-DNA insertion site	109
4.3.1	Rationale	109
4.3.2	Isolation and characterization of wildtype genomic clones	109
4.3.3	Restriction map of L-EN1-4	110
4.3.4	Sequencing the wildtype genome	112
4.4	Isolation of cDNAs mapping to the T-DNA tagged region	113
4.5	Summary	115

Chapter 5: Two overlapping genes in the tagged region

5.1	Introduction	116
5.2	Sequencing the cDNAs	117
5.2.1	Partial sequencing reveals similarity to the tagged genome	117
5.2.2	The cDNAs correspond to two distinct but highly related transcripts	117

5.3	Overlapping of the <i>PKT1</i> and <i>PKT2</i> transcription units	119
5.3.1	Rationale	119
5.3.2	Restriction map of the <i>PKT1</i> gene in the wildtype genome	120
5.3.3	Construction of a restriction enzyme map for pENB8	121
5.3.4	Sequencing the E5.1E5.2 genomic fragment	121
5.3.5	Searching for additional exons of <i>PKT2</i>	121
5.4	Approaching the 5' ends of the <i>PKT1</i> and <i>PKT2</i> transcripts by 5' RACE	122
5.4.1	Rationale	122
5.4.2	Applying the SLIC method to the <i>PKT1</i> and <i>PKT2</i> transcripts	123
5.4.2.1	Isolation of mRNA	123
5.4.2.2	SLIC of <i>PKT1</i> and <i>PKT2</i> transcripts	123
5.4.2.3	SLIC of the control human transferritin RNA	124
5.4.3	Development of a new 5' RACE system (LOPRACE)	124
5.4.4	RACE by homopolymeric tailing	125
5.5	Search for ESTs of the <i>PKT1</i> and <i>PKT2</i> genes	127
5.6	Summary	128

Chapter 6: Complex organization of the T-DNA region in AtEN101 129

6.1	Introduction	129
6.2	5'RACE of fusion transcripts	129
6.2.1	Application of the SLIC method	129
6.2.2	Sequencing of SLIC products	131
6.2.3	5' RACE by homopolymeric tailing	131
6.2.4	Fusion transcripts with unexpected 5'ends	131
6.3	Organization of T-DNA in AtEN101	133
6.3.1	Rationale	133
6.3.2	Southern blot detection of <i>gus</i> genes in AtEN101	133

6.3.3	Restriction enzyme analysis of the T-DNA junctions	135
6.3.3.1	PCR amplification of the left and right junctions	135
6.3.3.2	Restriction analysis of the left junction	136
6.3.3.3	Restriction analysis of the right junction	137
6.3.3.4	Sequences of the junction segments	138
6.4	Summary	138

Chapter 7: Identity of the *PKT1* and *PKT2* genes 140

7.1	Introduction	140
7.2	Identification of Open Reading Frames in the cDNA sequences	141
7.3	Search for sequences with similarity to the ORFs	143
7.4	Transcripts mP18 and mP19 encode 3-ketoacyl-CoA thiolases	143
7.5	Conserved primary structure features in P18-1 and P19-1	145
7.6	Prediction of secondary structure features	148
7.7	Construction of a 3D model for P19-1/P18-1	151
7.8	Cellular localization of the putative P18-1 and P19-1 thiolases	152
7.8.1	P19-1 is a putative peroxisomal protein (PKT2)	152
7.8.2	Absence of targeting signals in P18-1 (PKT1)	153
7.8.3	Determining the localization of PKT1 and PKT2	154
7.9	Development of a new and more powerful signature for thiolases	157
7.10	Summary	158

Chapter 8: The promoter regions of *PKT1* and *PKT2* 161

8.1	Introduction	161
8.2	Sequencing the putative promoter region of <i>PKT1</i>	162
8.2.1	Generation of Tn1000 insertion-derivatives of pENB8	162
8.2.2	Selection of pTB8-x clones	163
8.2.3	The sequence of the <i>PKT1</i> promoter region	164
8.3	Sizing the fusion transcripts by Northern analysis	167
8.4	The promoter region of <i>PKT2</i>	168
8.4.1	Searching for <i>PKT2</i> transcripts in AtEN101	168
8.4.2	Matches to <i>cis</i> -regulatory elements in the promoter region of <i>PKT2</i>	168
8.5	Functional analysis of the putative promoter regions	169
8.5.1	Rationale	169
8.5.2	Construction of the test plasmids	169
8.5.2.1	Putative promoter region of <i>PKT1</i>	169
8.5.2.2	Test plasmids from the putative promoter region of <i>PKT2</i>	172
8.5.3	Transient expression assays	172
8.5.3.1	Fundamentals of the assay method	172
8.5.3.2	Transient assays	173
8.6	Summary	174

Chapter 9: Expression of *PKT1* and *PKT2* and the application of AtEN101 as a molecular marker 178

9.1	Introduction	178
9.2	Transcriptional expression of <i>PKT1</i> and <i>PKT2</i>	179
9.2.1	Northern blot analysis of transcription	179

9.2.2	Specific detection of <i>PKT1</i> and of <i>PKT2</i> transcripts by RT-PCR	180
9.2.3	Differential expression of <i>PKT1</i> and <i>PKT2</i> transcripts	181
9.2.4	Semi-quantification of <i>PKT1</i> and <i>PKT2</i> transcripts	182
9.3	Studies on the regulation of <i>gus</i> expression	182
9.3.1	Environmental stresses	183
9.3.2	Effects of growth substances	183
9.4	Expression of the marker in mutant backgrounds	185
9.5	Summary	187

Chapter 10: A family of genes for 3-ketoacyl-CoA thiolases in *A. thaliana* 187

10.1	Introduction	187
10.2	The structure of the <i>PKT1</i> and <i>PKT2</i> genes	188
10.2.1	Restriction map of pENB3	188
10.2.2	Sequencing of pENB3	189
10.3	Genes encoding isozymes of PKT1 and PKT2	190
10.3.1	Searching for additional genes by Southern blot analysis	190
10.3.2	Identification of ESTs of <i>PKT</i> genes	192
10.3.3	A fourth, putative gene for peroxisomal 3-ketoacyl-CoA thiolase	193
10.4	Comparative analysis of the <i>PKT</i> genes	193
10.4.1	The exon/intron structure of the <i>PKT</i> genes	193
10.4.2	Relatedness of the <i>PKT</i> genes	193
10.5	Summary	196

Figure 4.5 Strategy for sequencing the pIPCR2 insert and distribution of stop codons.

- A)** Sequencing strategy. The various organizational elements of the insert are indicated as in Figure 4.4B. Restriction enzyme sites used to generate subclones are indicated: A1: *Apa* I; B1: *Bam*H I; E1: *Eco*R I; E5: *Eco*R V. The fragments subcloned are shown as horizontal bars aligned with their corresponding region on pIPCR2. The names of the subclones and primers used for sequencing are also indicated. Arrows show the lengths and positions of the determined sequences. Open arrowheads indicate the position and orientation of specific primers used in sequencing.
- B)** Distribution of stop codons on the native plant genome portion of the sequence and the downstream flanking T-DNA up to -689 (-1 was arbitrarily assigned to the first position of the T-DNA). A composite of the sequenced junction region downstream from the *Bam*H I site and the non-sequence but presumed intact segment of the LB up to -689 was submitted to the program FRAMES of the GCG package (Devereux *et al.*, 1994). The graphical output of the program is shown. The horizontal bars represent the sequence in the six-reading frames. The short vertical bars above and below the lines indicate the occurrence of methionine and stop codons, respectively. Assuming intactness of the T-DNA downstream from the sequenced region, the reporter gene *gus* is in frame 2.

Chapter 11: Discussion	197
11.1 Overview of the results	197
11.1.1 The transgenic line AtEN101	197
11.1.2 Isolation of a new morphogenetic mutant	198
11.2 pΔ <i>gus</i> Bin19 as a promoter trap	199
11.2.1 Cloning genes with promoter trapping	199
11.2.2 Validity of promoter traps as functional markers	199
11.2.3 Promoter trapping in <i>Arabidopsis</i> with pΔ <i>gus</i> Bin19	202
11.2.4 The use of <i>gus</i> as a reporter gene	204
11.2.5 Final considerations	204
11.3 Two overlapping <i>PKT</i> genes in <i>A. thaliana</i>	205
11.3.1 Unusual genic organization of <i>PKT1</i> and <i>PKT2</i>	205
11.3.2 The promoter regions of <i>PKT1</i> and <i>PKT2</i>	206
11.4 <i>PKT1</i> and <i>PKT2</i> as peroxisomal thiolases	208
11.4.1 Are <i>PKT1</i> and <i>PKT2</i> real proteins?	208
11.4.2 Intracellular targeting of <i>PKT1</i>	209
11.4.3 The peroxisomal targeting motif PTS2	210
11.4.4 A family of <i>PKT</i> genes	210
11.4.5 The specific functions of <i>PKT1</i> and <i>PKT2</i>	211
11.4.6 New signature sequences for thiolases	213
11.5 AtEN101 as a functional marker	214
11.5.1 GUS activity and <i>PKT1</i> / <i>PKT2</i> expression	214
11.5.2 GUS activity in embryo-defective mutants	215
11.6 <i>bashful</i> : a new shape mutant	216
11.6.1 Non-genetic nature of morphological aberrations in <i>bashful</i> embryos	216
11.6.2 The nature of the affected function in <i>bashful</i>	216
11.6.3 A strategy to isolate <i>BASHFUL</i>	217
11.7 Future work	217
11.7.1 <i>bashful</i>	218
11.7.2 pΔ <i>gus</i> Bin19	218

11.7.3	PKT genes	218
Appendix 1		
	Alignment of insert in pIPCR2 with genomic B1.4-E1.5 fragment	220
Appendix 2		
	Chromatogram of 5' RACE plasmid pTM-7	222
Appendix 3		
	Alignment of PLFR-7 and p Δ <i>gus</i> Bin19 sequences	223
Appendix 4		
	List of deoxyoligonucleotides	224
Appendix 5		
	Alignment of EST T75842 with the mP18 sequence	225
Appendix 6		
	Alignment of EST AI100014 with the mP19 sequence	226
Appendix 7		
	Multiple alignment of ESTs from putative <i>PKT5</i> transcripts	227
Appendix 8		
	Alignment of the <i>PKT2</i> , <i>PKT2</i> , and <i>PKT3</i> sequences	228
Bibliography		234

Abbreviations

2,4 D	2,4-Dichlorophenoxyacetic acid
2,4 DB	2,4-Dichlorophenoxybutyric acid
5' RACE	Rapid amplification of cDNA ends
A	Optical absorbance
ABA	Absciscic acid
ABRE	Absciscic acid responsive element
AFLP	Amplified fragment length polymorphisms
BAC	Bacterial artificial chromosome
bp	base pair(s)
BSA	Bovine serum albumin
CAPS	Cleaved amplified polymorphic sequence
cDNA	complementary DNA
CoA	Coenzyme A
CTAB	Cetyltrimethylammonium bromide
DAF	Days after flowering
DEPC	Diethylpyrocarbonate
DMSO	Dimethylsulfoxide
DNAase	Deoxyribonuclease
dNTP	deoxynucleoside triphosphate
DTT	Dithiothreitol
EDTA	N,N'-Ethylenediaminetetracetic acid
EST	Expressed sequence tag
g	relative centrifugal force
GA ₃	Gibberellic acid
HAF	Hours after flowering
HEPES	N-2-hydroxyethylpiperazine-N'-2-
ethanesulphonic acid	
IAA	Indol-3-ylacetic acid
IPTG	Isopropyl-β-D-thiogalactopyranoside
IPTG	Isopropylthio-β-D-galactoside
kb	kilobase(s)
LB	Left border of T-DNA
LOPRACE	Loop-primed rapid amplification of cDNA ends

M	Molar
MES	Morpholineethansulfonic acid
min	minute
MOPS	3-(N-morpholino)-propanesulfonic acid
N	Normal
nt	nucleotide(s)
O/N	overnight
ORF	Open reading frame
PCR	Polymerase chain reaction
PEG	Polyethyleneglycol
pfu	plaque forming units
PTS1	Peroxisomal targeting signal 1
PTS2	Peroxisomal targeting signal 2
RACE	Rapid amplification of cDNA ends
RAPD	Random amplified polymorphic DNA
rDNA	Ribosomal DNA
RFLP	Restriction fragment length polymorphism
RNAase	Ribonuclease
rpm	rotation per minute
RT	Room temperature
RT-PCR	Reverse transcription PCR
SDS	Sodium dodecylsulfate
secs	seconds
Sevag	Mixture of chloroform/isoamyl alcohol (24:1)
SSC	Sodium chloride/Sodium citrate
SSLP	Single sequence length polymorphism
SURE	Sucrose responsive element
TE	Tris (10 mM), EDTA (1 mM) solution
TEMED	N,N,N',N'-Tetramethylethylenediamine
TIS	Transcription initiation site
Tris	2-Amino-2-(hydroxymethyl)-1,3-propanediol
UTR	Untranslated region
v/v	volume per volume
vol	volume
w/v	weight per volume
X-gal	5-bromo-3-indolyl- β -D-galactoside
X-Gluc	5-bromo-4-chloro-3-indolyl β -D-glucuronic acid
YAC	Yeast artificial chromosome

Chapter 1

Introduction

1.1 An historical perspective on the study of development

"There is considerable difficulty in understanding how the plant is formed out of any seed or any animal out the semen. Everything that comes into being or is made must be made out of something, be made by the agency of something, and must become something"

Aristotle, *in* "Generation of Animals" 733b20

The predictable, heritable transformations undergone by an individual organism throughout its life cycle have long been a source of wonder and object of enquiry of Humanity. Among the chief questions pertaining to the subject are those concerning the causal events of ontogeny and the mechanisms by which it is attained. The pursuit of an understanding of development and its interpretation have been shaped throughout the ages by different philosophical and technological contexts. Here, the major steps and movements that have characterized this quest are reviewed to situate the present scientific approaches and models of development in the historical continuum that shaped them.

In addition to the bibliographic sources specifically used for and mentioned in parts of the text, the writing of this Section is indebted to (Gilbert, 1991; Hamburgh, 1971; Portugal & Cohen, 1979; Slack, 1983; Stubbe, 1965).

1.1.1 Beginnings

Of the Antiquity authors, Aristotle was foremost among those concerned with the subject. He was interested not only in the nature of the living creatures but by what means they came into being:

"Now that out of which it [living being] is made is the material...but we now are inquiring not out of what the parts of an animal are made but by what agency".

ibid. 733b

Aristotle's method was dialectical and based on careful observation of the natural world. His thought and conclusions were to pattern much of the western view of the world until the Renaissance. The ability to look at the World on a different scale offered by the invention of the microscope in the XVI century was pivotal to the enormous progress that followed in biological

studies. In particular, the introduction of the compound microscope in the XVII century enabled the multiplication of detailed morphological studies (reviewed in Churchill, 1994). A favourite subject since Aristotle, the development of the chicken embryo was described from its early stages by prominent microscopists including Marcello Malpighi and later, in the XVIII century, Albrecht Von Haller, Lazzaro Spallanzani and Friedrich Wolff (1733-1794). Malpighi derived from his observations of the chick embryo a belief on preformation as the basis of development (Stubbe, 1965). This found support among his contemporaries including Von Haller and Spallanzani. By the contrary, Wolff held the epigenetic view that an essential force, "*vis essentialis*", was required for growth and development of organisms from undifferentiated cells (ampullae). A purely descriptive character of the investigations was to be the hallmark of developmental studies throughout most of the XIX century (Churchill, 1994). During this period technological improvements of the microscope, the use of the microtome and advances in the techniques of preparation of specimens all contributed to a great increment in the number of studies and of students of development.

The XIX century was a revolutionary time for the biological sciences, proliferous in new theories and modes of enquiry. Of particular impact on the study of development were the cell theory of Schleiden and Schwann presented in 1839, itself an embryological theory, and its modifications with the recognition by Rudolph Virchow in 1858 that "*Omne cellula e cellula*". Broad concepts of transmutation of living forms were also introduced, initially with Lamarck's contribution in 1800 and culminating with the joint presentation of the theory of evolution of Wallace and Darwin in 1858. In his book "*Origins of species*" Darwin, recognizing the parallel between the degree of similarity of the early ontogenic stages in different species and their evolutionary relatedness, presented the law of embryological resemblance suggesting that it could be the basis for evolutionary understanding. An extreme view based on this relation was defended by Ernst Haeckel with his theory of recapitulation of phylogeny during ontogeny (Churchill, 1994). This view had the merit of greatly increasing interest in embryology albeit for the sake of studying evolution. Of particular significance was the establishment of comparative embryology in the first half of the century by Karl Ernst von Baer who used it to derive general principles of development from empirical observation. Subjacent to it was the notion of homology.

Despite their significant increase, studies on development throughout most of the XIX century were based on simple observation of the natural world, the same approach used by Aristotle to derive his conclusions. A turning point in the investigation of ontogenesis occurred towards the end of the century with the adoption of experimental manipulation of the subject of study (reviewed in Maienschein, 1994). Among the contributors to this new approach Wilhelm Roux is the best known. In his most distinguished work published in 1888, Roux set out to test the hypothesis that the differentiation in the embryo was due to internal factors rather than to external conditions. He did so by killing one of the cells in two-cell stage frog embryos and

observing the fate of the one remaining. The live cell continued to develop and produced a partial embryo, a result that was interpreted by Roux as indicative that the different parts of the early embryo developed independently of each other.

1.1.2 Development and Heredity

The application of experimental testing would have far reaching consequences for the unravelling of the links between ontogenesis and heredity. The intimate relation between these two concepts was well known to Aristotle who enunciated:

"For a given seed does not give rise to any chance living being, nor spring from any chance one; but each springs from a definitive parent. And thus it is that from which the seed comes which is the origin and fabricator of its offspring"

In "Generation of Animals" 645b25

That both the male and female parents contributed to their offspring was accepted from the great philosopher's time. Tangible, direct evidence of it, however, was only obtained in 1846 with the viewing by Amaci of the fusion of the pollen tube with the egg cell in orchids. Later studies, in the period between 1875-1880, established that in both animals and plants fertilization was accompanied by fusion of the nuclei of the egg cell and the sperm or pollen grain (reviewed in Churchill, 1994). Shortly after, in 1885, Weissman presented his germ-plasm theory of heredity. It stated "that heredity is brought about by the transference from one generation to another, of a substance [the germ-plasm] with a definite chemical, and above all, molecular constitution" (cited in Whitehouse, 1973). At the time, the replicating character of the cell nucleus recognized by the botanist Strasburger in 1879 was taken as evidence that it was the carrier of hereditary information, as was accepted by Weissman. Strong support of this view came from investigations in experimental embryology. Boveri (1889) fragmented sea urchin eggs by shaking and found that enucleated pieces could sometimes be fertilized. These fragments developed into normal, albeit smaller, larvae. In other experiments involving two species of sea urchins Boveri found that cross-fertilization resulted in larvae with intermediary characteristics to those typical of the parent species. However, fertilization of enucleated eggs of one species with sperm from the other species yielded larvae identical to those derived from the male parent species, although smaller in size.

The re-discovery of Gregor Mendel's work in 1900 led to a search for the physical support for the inherited characters or genes. By the turn of the century, a substantial body of evidence had accumulated on the behaviour of chromosomes which, as recognized by many, was compatible with their being carriers of genetic information (reviewed in Sturtevant, 1965). Boveri's experiments with sea urchins eggs were also central to establishing the chromosomes as the carriers of genetic information. It was known that if the four cells resulting from the first

two cleavages of a fertilized egg are separated each would develop into a normal embryo. In 1902 Boveri showed that following polyspermic (double) fertilization of eggs, the 3 haploid chromosome sets would be distributed more or less randomly by the 3 or 4 cells resulting from the first cleavage. When separated, some, but never all of the cells derived from the same egg, would develop into normal embryos. He demonstrated that the normal development of a cell required the presence of a complete haploid set of chromosomes and that each chromosome contributed differently to the process. The following years saw the construction of a chromosome theory of inheritance which became well established in 1915 with the publication of "The Mechanism of Mendelian Heredity" by Thomas Morgan, Sturtevant, H. Muller and C. Bridges.

In spite of the contributions of developmental studies to the field of genetics, the application of the latter to ontogenic research was not immediate. The discipline of developmental genetics had a prolonged birth originating from interest in unexpected non-Mendelian ratios of genetic inheritance (reviewed in Gilbert, 1994; Wilkins, 1986). In many instances the ratio deviations were found to be caused by recessive embryo-lethal mutations. It was only in 1938 that the new area of research was explicitly framed by Salome Gluecksohn-Schoenheimer: "A mutation that causes a certain malformation as the result of a developmental disturbance carries out an 'experiment' in the embryo by interfering with the normal development at a certain point. By studying the details of the disturbed development it may be possible to learn something about the results of the 'experiment' carried out by the gene" (cited in Gilbert, 1994). During its initial period developmental genetics aimed at identifying and characterizing the earliest stages of mutant phenotypes (the phenocritical stage) with the aim of inferring the function of the mutated gene (Gilbert, 1994; Wilkins, 1986). The major contribution of these studies to the understanding of development was the recognition that the process is paralleled by a spatially and temporally organized programme of gene expression.

Following the identification of chromosomes as the carriers of genetic information, the field of heredity became focused on determining the nature of the gene and its mechanisms of action. The findings prompted by these quests were to revolutionize biology, create the disciplines of molecular biology and molecular genetics, and become seminal of a new understanding of development. The key steps in this revolution were the demonstration of DNA as the genetic material in cells by Avery, MacLeod and McCarty in 1944 and the resolution of its double-stranded chain structure by Watson and Crick in 1953 (Sturtevant, 1965), and the advancement of the operon model by François Jacob and Jacques Monod in 1961 (Wilkins, 1986). The latter model enshrined the notion of control of development by regulation of gene activity. It established and was based on the concept of regulatory genes, and of a unstable molecular intermediate, the mRNA, between the gene and its ultimate (protein) product of expression.

Even though the operon model was derived from prokaryotes, it was quickly adopted and adapted to explain regulation of developmental change in higher organisms (reviewed in Wilkins, 1986). Furthermore, it was recognized that the control of gene expression could occur at several stages of the process from transcription to post-translationally. This vision of development as a complex process of change hierarchically controlled and modified largely through regulation of gene expression is currently dominant (*e.g.* Lindsey & Topping, 1993). In this scientific context, the identification of the actors involved, *i.e.* the genes and their expressed activities, and a mechanistic understanding of their action and its regulation are essential to the construction of models of development.

1.2 Embryogenesis in higher plants

1.2.1 Generalities

The transformation of a single cell, the zygote, into a complex, well-organized multicellular embryo, is one of the most fascinating developmental phenomena. In plants, embryogenesis is initiated by the fertilization of the egg cell by a sperm cell continuing until the new sporophyte is competent and ready to establish itself as an independent organism. This competence in higher plants is achieved by the laying of the basic body plan of the seedling including the specification of the apical meristems required for post-germinative growth, and the precursors of the main tissue types. It also involves the generation of storage organs on which the seedling relies during germination. Usually, the maturation of the embryo and the seed is followed by the onset of dormancy, *i.e.* a stage of virtual arrest of metabolic activity, growth and development accompanied by desiccation, that is an adaptation to optimize the time of germination.

A unique feature in the sexual reproduction of angiosperms is that in parallel with the formation of the zygote another fertilization takes place, that of the embryo-sac by the second germinative sperm cell from the pollen grain. Fusion of their nuclei creates a triploid cell that evolves into the endosperm, a tissue that envelops and is thought to be a source of nutrients to the developing embryo (Lindsey & Topping, 1993; Steeves & Sussex, 1989). This nutritive tissue may become a storage tissue in the seed, although in many dicots it is absorbed during seed development (Steeves & Sussex, 1989). In addition to its nutritive role the endosperm seems to contribute an environment satisfying the requirements of embryogeny. *In vitro* culture of excised embryos has shown that the high osmolarity found in the endosperm is apparently one of the factors involved, particularly at earlier stages of embryogenesis (Raghavan, 1986; Steeves & Sussex, 1989). The presence of hormones or other biochemical factors seems to be another.

A relevant aspect for the study of plant embryogeny is that the zygote is not unique in its capacity to undergo this transforming process. Other types of cells, both of the sporophytic

and gametophytic generations, will develop into embryos upon cultivation in appropriate media (Steeves & Sussex, 1989). Some plants are known where embryos are naturally formed on mature organs. Documented examples include the orchid *Malaxus paludosa* where terminally-differentiated leaf cells can become meristematic and form embryos, and, similarly, the formation of somatic embryos that occurs at the surface of the stem in *Ranunculus sceleratus* (Steeves & Sussex, 1989). A more widespread phenomenon is apomixis where embryos derive from sporophytic tissues of the ovule or directly from an embryo sac (reviewed in Koltunow *et al.*, 1995). Usually, for a given species, the somatic embryos obtained by *in vitro* culture are structurally distinct from their zygotic counterparts at the various stages, but both types share the same organization and developmental timing (Zimmerman, 1993).

Many of the difficulties associated with the study of zygotic embryogenesis derive from the relative inaccessibility of the ovules and embryos within the developing angiosperm fruit. This barrier is absent when somatic embryogenesis takes place *in vitro* or excised zygotic embryos are grown in culture. Furthermore, the composition of the culture media can be closely controlled and manipulated allowing for the investigation of its influence in embryogeny and the study of this process in a well defined environment. These possibilities led to the adoption of *in vitro* somatic embryogenesis as a system for the investigation of plant embryogeny (Raghavan, 1986; Zimmerman, 1993). The extent to which work with these systems is relevant to zygotic embryogenesis is determined by the overlap between the two processes. Whereas body patterning, morphogenesis and gene expression programs share an overall similarity in somatic and zygotic embryogenesis (Dodeman *et al.*, 1997; Zimmerman, 1993), it is yet largely unknown if they are controlled by the same regulatory networks.

Although in angiosperms zygotic embryogenesis is usually a well defined period clearly delimited by dormancy, it is nevertheless morphogenically continuous with the subsequent development of the plant. Not only are the meristems and the storage organs on which germination and post-germinative growth depend established in the forming embryo, as dormancy is not required for morphogenic developmental continuity. This is demonstrated by vivipary, *i.e.* the absence of a seed dormancy period, in the life cycle of some angiosperms, as well as other groups of vascular plants. Mangroves, *e.g.* *Rhizophora* and *Bragiera*, provide perhaps the best known example of viviparous species (Skene, 1939). Vivipary has also been documented in mutants of several species where normally a dormancy period exists. These include the *aba-1* mutant of *Arabidopsis thaliana* characterized by germination of the seeds while in silique in a humid atmosphere (Koornneef *et al.*, 1982) and the *viviparous* mutants of maize (McCarty & Carson, 1991). These mutants may be affected in other aspects of the embryogenic programme: in the several alleles of *viviparous-1* the maturation phase is altered and several genes normally expressed during this stage fail to do so (McCarty & Carson, 1991). Precocious germination of embryos can also be experimentally induced by their excision from the seed prior to the onset of dormancy and *in vitro* cultivation (in the absence of hormones; *e.g.*

Bisgrove *et al.*, 1995). Evidence for developmental continuity is also found in the wide range of morphological complexity attained by embryos of different species upon maturation. Whereas in *A. thaliana* the mature embryo consists simply of a root and an apical meristems, a hypocotyl and two cotyledons (Lindsey & Topping, 1993), in maize, several leaves and adventitious roots will have been initiated, and the primary root has started being formed (Steeves & Sussex, 1989).

The morphogenic history of the mature embryo is species-specific. In common, however, is the establishment of a recognizable distribution of cells, tissues and organs, namely pattern formation (reviewed in *e.g.* Lindsey & Topping, 1993). Also, in spite of the extent of morphological development of a mature embryo varying substantially with species, its basic features are remarkably similar among vascular plants (Raghavan, 1986; Steeves & Sussex, 1989). The main common trait is the laying down of a polar axis whose extremities are defined at maturity by the shoot and the root apical meristematic regions. In angiosperms, the polar axis has cytological radial symmetry throughout most of its length, except at the sites of attachment of the cotyledon(s), a distinct organ system. Determining how this basic body plan is laid down, understanding the mechanisms of cell division and cell differentiation and the regulatory circuitries that control these processes, as well as the relationships that exist between them is fundamental to a global comprehension of angiosperm embryogeny.

1.2.2 Arabidopsis as a model biological system

The study of many biological phenomena has often focused on a restricted number of species thought to be representative of a much wider group of organisms. The sharing of basic morphogenetic patterns among angiosperms, a likely consequence of their evolutionary relatedness, is certain to reflect a similitude of the underlying developmental processes and regulatory mechanisms. This observation justifies the concentrated study of the various aspects of embryo formation in a restricted number of species chosen as models. Historically, however, in plants, whereas genetic studies have focused on crop species such as maize, wheat and rice because of their enormous economic importance, developmental work was dispersed by a wide number of systems (*e.g.* Raghavan, 1986). The advent of the molecular era has changed this scenario. The availability of tools to clone and analyze genes led to the search for model plants that would facilitate these tasks. One plant that proved to be ideally suited to these purposes was *A. thaliana*.

The characteristics that made the crucifer *A. thaliana* a model biological system were not limited to its small size and rapid cycling, the ability to self-pollinate and to be cross-pollinated in a controlled manner, and its high seed set, all ideal for classical genetics (Meyerowitz, 1987; 1992). Principally, they were related to its nuclear genome. *Arabidopsis* has one of the smallest known genomes among flowering plants, estimated to be about 100 Mb (Meyerowitz, 1987; 1992), which contrasts with the much larger genomes of other species,

including the commonly used tobacco (1,600 Mb), pea (4,500 Mb), and wheat (5,900 Mb), that can range up to 12,700 Mb in *Fritillaria assyriaca* (Bennett & Smith, 1976). Importantly, the repeated DNA fraction in *Arabidopsis* is limited to about 15%, with little of it, estimated at 1%, occurring in an interspersed pattern (Meyerowitz, 1987; 1992). This is also exceptional among the genomes of higher plants where large proportions of repeated DNA sequences are the norm (Flavell, 1980). These attributes have made *Arabidopsis* ideal for molecular genetics studies. Unlike in many other species, the construction of physical maps of the *Arabidopsis* genome is feasible with current technology and chromosomal walking possible allowing for positional cloning of genes of interest (see Section 1.3.4.3). These advantageous characteristics of *Arabidopsis* together with the strong evolutionary ties among angiosperms have caused it to be elected as a model system also for developmental studies of this group of vascular plants, particularly for dicots (*e.g.* Lindsey & Topping, 1993). For monocots, the large body of knowledge accumulated on maize has made it a model system together with rice which benefits from the smallest known monocot genome.

In the last several years an international effort has been undertaken to completely sequence the genome of *Arabidopsis* and identify its genes (Multinational Coordinated *Arabidopsis thaliana* Research Project, 1997). A large scale project to identify transcription units by partial sequencing of cDNA clones has already generated over 38,000 Expressed Sequence Tags (ESTs; Meinke *et al.*, 1998). Representative genomic libraries that allow for physical mapping have been constructed in a variety of vectors including Yeast Artificial Chromosomes (YACs), Bacterial Artificial Chromosome (BAC), and P1 (Liu *et al.*, 1995; Meinke *et al.*, 1998). Completion of genome sequencing, expected for the year 2000 (Meinke *et al.*, 1998), will permit the physical mapping of the transcription units represented in the ESTs. Genetic analysis has been facilitated by the construction of recombinant inbred lines (Lister & Dean, 1993). By resorting to these lines a linkage map of over 800 molecular markers including Restriction Fragment Length Polymorphisms (RFLPs), Single Sequence Length Polymorphisms (SSLPs), Cleaved Amplified Polymorphic Sequences (CAPSs), ESTs, cloned genes and ends of BAC and YAC artificial chromosomes has been constructed (Meinke *et al.*, 1998). A classical genetic map with over 100 genes is also available (Meinke *et al.*, 1998). In addition, *Arabidopsis* is susceptible to *Agrobacterium* infection which has permitted the use of molecular tagging techniques to explore its genome (See Section 1.3.5), for clonal analysis (*e.g.* Scheres *et al.*, 1994), and for functional studies. *Arabidopsis* is thus in an enviable position for the identification, cloning and characterization of genes of interest.

1.2.3 Zygotic embryogenesis in *A. thaliana*

In *Arabidopsis* the egg cell is found in ovules of the anatropous type located along the length of the siliques. Ontogenesis of the embryo-sac is of the *polygonum*-type, the same as that of maize and of about 70% of the species where it has been investigated (reviewed in Reiser & Fischer, 1993). In this type of monosporic development only the megaspore at the chalazal end will become functional of the four formed by meiosis of the megasporocyte. The others degenerate and are crushed by the growing megaspore. During megagametogenesis the megasporocyte undergoes two successive rounds of mitoses originating eight nuclei and seven cells, the central cell being binuclear. The egg cell flanked by two other cells, the synergids, is located at the micropylar end of the ovule.

Prior to fertilization the pollen tube grows through the micropyle triggering a process involving one of the synergids that results in the fusion of a sperm nucleus with that of the egg cell. The other sperm nucleus and the two nuclei of the central cell fuse together originating the precursor of a triploid endosperm (reviewed in Lindsey & Topping, 1993). A noticeable feature of the newly formed zygote is its structural polarity. In the newly formed cell a large number of small vacuoles are found at the micropylar end and most cytoplasm is concentrated at the chalazal end (Mansfield & Briarty, 1991). The vacuoles eventually coalesce forming a large one located towards the micropylar end. This axial polarization is preceded by structural polarity of the egg cell and that subjacent to the events that form the embryo-sac.

The development of the embryo has been described in detail for *Arabidopsis* and reviewed previously (Dolan *et al.*, 1993; Lindsey & Topping, 1993; Mansfield & Briarty, 1991; West & Harada, 1993). The patterns of cell division, positioning and differentiation are closely controlled, particularly during the early stages. Initiation of embryo formation involves elongation of the zygote along the micropylar-chalazal axis normally followed by a transversal asymmetric cell division (Mansfield & Briarty, 1991). The cell nearer the micropyle, the basal cell, retains a highly vacuolated character and will give rise to the suspensor. It will also contribute to the root meristem quiescent centre and the central root cap of the embryo. The smaller spherical and cytoplasmically-rich terminal cell at the chalazal end will develop into the remainder of the embryo proper. Within the first 24 hours after flowering (HAF) the terminal cell divides longitudinally twice forming a quadrant embryo, and at 30 HAF their transverse division results in an octant embryo. At the octant stage an apical and a basal tier of four cells each can be distinguished separated by the so-called O' boundary, each tier contributing to different parts of the mature embryo (Lindsey & Topping, 1993). Periclinal division of each of the eight cells originates a 16-celled embryo featuring a structurally distinct outer layer, the protoderm. The early globular stage is reached by further division of these cells. At the same time, the suspensor has completed its development consisting of a large basal cell and a filament of six to eight cells. Of these, only the one at the apical position, the hypophysis, contributes to the embryo.

A morphologically significant transition, the end of the globular stage, is reached about 3 days after flowering (DAF) when an increase in mitotic activity with bilateral symmetry is observed in the areas that will originate the two cotyledons (Mansfield & Briarty, 1991). The increased cell division results in a short-lived triangular shaped-embryo that progresses into a heart shape. Vascular tissue begins to differentiate at the base of the cotyledons and plastids begin to increase in number (Mansfield & Briarty, 1991) and the shoot meristem initials can be distinguished (Barton & Poethig, 1993). A columella and central cells of the radicle are evident at the late heart stage (Dolan *et al.*, 1993). By 4 DAF the elongated cotyledons confer a torpedo likeness to the embryo (Mansfield & Briarty, 1991). Also, the central cells of the hypocotyl and radicle divide forming the primary vascular tissue surrounded by three layers of parenchymatous cells.

Continual growth of the embryo is accompanied by further differentiation of tissues, and elongation within the limited space of the seed forces the cotyledons to curve throughout what are known as the walking-stick and the cotyledonary stages. On maturity the *Arabidopsis* embryo consists of about 20,000 cells distributed by the two cotyledons and a simple embryo axis where the root and shoot meristems have not originated any additional structures (Lindsey & Topping, 1993). Thus, along the apical-basal axis the pattern elements found include the root meristem and root, hypocotyl, cotyledons and shoot meristem (Jürgens *et al.*, 1991). Radial pattern formation results in the establishment of three concentric rings of tissue including the protoderm, ground tissue and vascular tissue (Dolan *et al.*, 1993; Jürgens *et al.*, 1991).

The body plan of the embryo is inherited by the juvenile germinated plant, *i.e.* the seedling, in *Arabidopsis*, as in flowering plants in general (see Section 1.2.1). The pattern elements distinguishable in the seedling can be traced to earlier stages of embryogenesis (Dolan *et al.*, 1993; Jürgens *et al.*, 1991). The upper tier of the octant stage proembryo develops into the shoot meristem and most of the cotyledons except the shoulder region. The lower tier of the octant stage proembryo contributes to the remainder of the cotyledons, the hypocotyl and most of the root structure. The quiescent centre and the central root cap derive from the hypophysis.

1.2.4 Gene activity and embryogenesis

1.2.4.1 Genic autonomy of the developing embryo

The morphogenetic changes that the developing embryo undergoes are accompanied by the expression of numerous genes both in the embryo and in the surrounding endosperm. Nucleic acid hybridization analysis indicates that the actively expressed genes at any particular stage range from about 12,000 in the wheat endosperm at 25 DAF, and 14,000 in soybean cotyledon stage embryo to 34,000 in mature embryos of cotton (Goldberg *et al.*, 1989). These results are indicative of embryogeny as largely dependent on the expression of the genetic complement of

the new sporophytic generation, as has generally been found for other multicellular organisms (see Section 1.1.2). Demonstration of their genetic independence for completion of embryogenesis comes from the successful culture *in vitro* of excised embryos. Proembryos of *Brassica juncea* of as early as 8-cell stage have been grown *in vitro* developing into normal, mature embryos (Liu *et al.*, 1993). Similar results have been obtained with heart stage and older embryos of *Arabidopsis* (Wu *et al.*, 1992), and with developing embryos of several other species (Raghavan, 1986). It can then be asked what is the nature of these genes, how is their expression regulated, and how do they contribute to the various aspects of embryogeny.

1.2.4.2 Approaches to identify embryonically active genes

Several approaches have been followed to identify the genes expressed in the developing embryo of higher plants. The high abundance of transcripts of storage protein genes made these among the earliest to be cloned (Thomas, 1993). For the identification of many other types of genes a genetic strategy based on the isolation of mutants where embryogeny is altered or disrupted has been very successful, particularly in *Arabidopsis* (Mayer *et al.*, 1991; Meinke, 1985) and maize (Sheridan & Clark, 1993).

In *Arabidopsis*, screening of siliques for mutations causing seed abortion or visible alterations in the embryo has resulted in the isolation of hundreds of embryo-lethal mutants (*emb*), and embryo-defectives of various types including *fusca* and pigmentation mutants (Feldmann, 1991; Meinke, 1985). On the basis of genetic analysis it has been estimated that over 10,000 genes are expressed in the *Arabidopsis* embryo during its formation, of which at least 500 are anticipated to be essential for successful completion of this morphogenetic programme (Franzmann *et al.*, 1995). Many of these genes are expected to encode proteins responsible for housekeeping functions (Franzmann *et al.*, 1995). For example, a mutant, *bio1*, was shown to be a biotin auxotroph that can be rescued by the addition of exogenous biotin (Schneider *et al.*, 1989). Two other embryo-lethal mutants have been found to have mutated nuclear genes for a plastid glycyl-tRNA synthetase (Uwer *et al.*, 1998) and a ribosomal protein of the same organelle (Tsugeki *et al.*, 1996).

Mayer *et al.* (1991) screened *Arabidopsis* seedlings of a mutagenized population to identify mutations that affect the morphogenetic programme but are not embryo-lethal. For this reason such mutants were expected less likely to be affected in housekeeping functions. This strategy was designed to identify genes affecting pattern formation in the embryo based on its developmental continuity with the seedling. In the *Arabidopsis* mutants identified in this manner morphological alterations present in the seedlings could indeed be traced to earlier defects in the embryo (Jürgens *et al.*, 1991; Mayer *et al.*, 1991).

For the genes affecting seedling morphology Jürgens and co-workers (Jürgens *et al.*, 1991; Mayer *et al.*, 1991) distinguished three classes of mutants: those with altered apical-basal pattern, those affected in the radial pattern, and a group where seedling morphology, but not

body pattern, was defective. The analysis showed that the apical-basal patterning was regulated independently of the radial pattern. In the mutant alleles of four genes (*gurke*, *fackel*, *monopteros*, and *gnom*) elements of the apical-basal pattern were absent or altered in the seedlings, however, the radial tissue pattern was intact. For example, in mutant alleles of *gnom* the root is absent and the cotyledons are either substantially reduced or not present (Mayer *et al.*, 1993). Intrallelic variation of phenotype was observed with the most severely affected seedlings being ball-shaped. Nevertheless all three tissue types including epidermis, ground and vascular tissues, are present and organized in a concentric, radial distribution in *gnom* seedlings. Apical-basal patterning elements also seem to be independently specified. Thus in *gurke* mutants only the apical region is affected with the cotyledons and the shoot meristem being absent, whereas only these elements are present in seedlings of *monopteros* alleles. In *fackel* mutants, the cotyledons are directly attached to the root of the seedling and the hypocotyl is missing. Two genes (*knolle* and *keule*) were found to affect the establishment of the radial pattern of the seedling, the apical-basal patterning also being altered in mutants of both genes (Jürgens *et al.*, 1991; Mayer *et al.*, 1991). Mutant seedlings of *keule* and *knolle* shared a lack of a normally formed epidermis. By contrast, the vascular and ground tissues were morphologically typical in *keule* whereas in *knolle* lumps of vascular tissue appeared to be embedded in abnormally bloated ground cells.

Study of embryo development in the various pattern mutants showed that the phenotypic defects in the seedlings had an embryonic origin. In the case of *gnom* the asymmetric cytokinesis of the zygote is affected, the two daughter cells being of about the same size (Mayer *et al.*, 1993). The apical cell and its descendants divide with abnormal and irregular patterns that are thought to account for the phenotypic variation observed in the seedlings. Although a suspensor is present, the division of the hypophysis is abnormal and a root primordium is absent in the mature embryo. Cotyledon primordia are also absent at the heart stage. Similarly, morphological defects were detected in heart stage embryos of the other pattern mutants (Mayer *et al.*, 1991).

The number of genes thought to be involved in pattern formation have been estimated to be about 15 to 50 (Jürgens *et al.*, 1991). This is a small fraction of the genes thought to be essential for completion of embryogenesis (see above). Many other genes involved in embryo formation are unlikely to be detected by either of the two genetic approaches discussed above. Included will be those encoding functions that are redundant or otherwise not essential for completion of the developmental programme under the growth conditions used. Identification of these genes requires distinct strategies. The isolation and characterization of mRNA populations from developing embryos offers the basis for several approaches to the identification of genes active in the process regardless of their redundancy (see Section 1.3.1). The use of promoter trapping and related techniques constitute another powerful set of tools allowing the

identification of transcription units and subsequent selection according to patterns of expression (see Section 1.3.5).

1.2.4.3 Regulatory complexities of the embryo gene expression program

Most of the genes for which mRNA transcripts are found in the developing zygotic embryo are also expressed during post-germinative growth in the seedling and the adult plant (Goldberg *et al.*, 1989). For others, however, corresponding steady-state mRNAs and protein products accumulate only at specific stages of embryogenesis. Their expression seems to be temporally co-ordinated with mRNA transcripts of sets of genes accumulating throughout embryogeny while others are found specifically in early embryogenesis, during the maturation phase, or during late embryogenesis (Lea). A distinct set of transcripts is restricted to both late embryogenesis and to early germination. Many of the genes are also spatially restricted in their expression as has been shown primarily by *in situ* hybridization techniques but, increasingly, by reporter gene technology (see Section 1.3.5.4). Spatial specificity can limit expression to specific cell types, regions or organs of the embryo. Among the genes active during embryogenesis those encoding storage proteins have been the best characterized as a class. Their expression is usually highly controlled being restricted to the embryo during the maturation phase and often to the storage areas like the cotyledons (reviewed in Goldberg *et al.*, 1989; Lindsey & Topping, 1993; Thomas, 1993). The patterns of accumulation of mRNA for cruciferins and napin in embryos of *B. napus* exemplify this spatial and temporal regulation of gene expression. Fernandez *et al.* (Fernandez *et al.*, 1991) showed by *in situ* hybridization that napin mRNAs accumulate in the embryo axis from the late heart stage, in the outer faces of the cotyledons during torpedo stage, and in the inner faces of the cotyledons during the cotyledonary stage. Cruciferin gene expression displayed a similar wave pattern but was delayed relative to that of napin by about 5 days. However, during these early stages, no storage protein mRNAs were found in the apical region derived from the upper tier of octant stage embryos, where the meristem will form. By contrast, during the maturation phase accumulation of napin mRNA was highest in the apical meristematic region. Cruciferin mRNAs were also more abundant in this region but were excluded from some areas of the meristem. Tight regulation of gene expression during embryogeny is also illustrated by the mRNA for the soybean Kunitz trypsin inhibitor, Kti3. Its transcripts were first detectable by *in situ* hybridization only at the basal end of the embryo from the late globular stage (Perez-Grau & Goldberg, 1989). This basal specification persisted and became restricted to the newly formed ground meristem cell layer from the heart stage. An example of a gene subject to a complex temporal-spatial regulatory pattern of expression is *ARABIDOPSIS THALIANA MERISTEM LAYER 1 (ATML1)*. Its activity was first observed in the apical daughter cell of the zygote and its progeny until the periclinal divisions of the octant stage embryo when the protoderm is formed (Lu *et al.*, 1996). In the resulting 16-celled globular embryo, *ATML1* transcripts are only

found in the protoderm cells for which the gene remains a marker until torpedo stage when transcription is no longer detectable. *ATML1* transcripts are again present in the mature embryo but restricted to the L1 layer of the shoot apical meristem. The *SHOOTMERISTEMLESS (STM)* gene activity is also detectable from an early stage of embryogeny, the 32-64 celled globular stage, when it is restricted to 1-2 cells in the apical region (Long *et al.*, 1996). By the late globular stage *STM* transcripts are present in an apical stripe of cells becoming limited to a small patch separating the points of attachment of the two cotyledons from the heart-stage.

Specificity of gene expression usually documented at the level of gene transcription has also been demonstrated at the level of protein accumulation. Earlier examples include the abundant storage proteins (reviewed in Lindsey & Topping, 1993; Thomas, 1993). Low abundance products also show spatial and temporal tissue-specific expression as is illustrated in somatic embryos of carrot by the target epitope of the JIM4 antibody raised against an arabinogalactan-protein (Stacey *et al.*, 1990). In this instance the epitope, J4e, was found in a group of cells located at the junction between the future root and shoot and also in a single well-defined layer at the surface of the embryo during heart-stage. At torpedo stage J4e is found in cells of two regions of the future stele and also cells associated with the cotyledonary provascular tissue.

The available evidence does show that the embryo is genetically autonomous in its development with many thousands of genes being expressed (Section 1.2.4.1) in a closely controlled and complex spatial-temporal programme. Unravelling the mechanisms of control of this program is one of the challenges of the study of embryogenesis.

1.2.4.4 Regulatory mechanisms of gene expression

Some of the most interesting questions pertaining to development concern the mechanisms that determine the fate of the cells. Is the identity of the cell decided by its genealogy or by its relative position in the multicellular body? Clearly, the identity of a cell is largely determined by its pattern of gene expression. The study of its regulation should shed light on the mechanisms controlling cell identity. Yet, very little is known about the regulatory mechanisms of gene expression in embryogeny even at the most studied level, that of transcription.

A role for the clonal history of a cell in determining its fate has been suggested (Fernandez *et al.*, 1991). This was prompted by the observation that the axial regional boundary found in the expression of storage protein mRNAs during various stages of embryo formation in *B. napus* was clonal in origin, traceable to the O' ring (see above, Section 1.2.4.3). The importance of the clonal origin of a cell in its differentiation is challenged by a variety of observations. In numerous species embryogenesis is characterized by irregularity of cell division (Raghavan, 1986), but pattern formation and cell differentiation are not affected. In *Arabidopsis*, where the various aspects of cell division are tightly controlled, clonal analysis with a *gus* reporter gene demonstrated that cell lineages of the central region of the heart stage

embryo were not limited in their fate (Scheres *et al.*, 1994). They could both contribute to the root as well as the hypocotyl, although a preference for certain regions was observed. By contrast, derivatives of the hypophysis were limited to the columella and the central region of the root.

There is considerable evidence supporting a predominant role for positional information in a variety of plant developmental processes (reviewed in Dawe & Freeling, 1991), and, increasingly in recent years, during embryogenesis. Recently, direct proof of inductive signalling in plants was obtained by laser ablation of cells in the root meristem of *Arabidopsis* (van den Berg *et al.*, 1995). It was observed that the dead cells were replaced with others by division of neighbouring cells, and that their correct differentiation was dependent on contact with more mature cells. Evidence for perception of position on cell differentiation in embryogeny was obtained from surgical transection of carrot somatic embryos at torpedo stage (Schiavone & Racusen, 1991). The transected embryos regenerated the root pole in a position dependent, proximal-distal replacement pattern. These results were interpreted as indicative of a capacity of the cells near the surgical margin to sense their location in the original structure and the absence of the severed portion, differentiating accordingly. Analysis of several mutations affecting embryo pattern formation has also been useful in the study of the mechanisms determining cell fate in developing embryos. For example, in *fackel* mutants the central region of the embryo which specifies most of the root and the hypocotyl in wildtype, is affected (see Section 1.2.4.2; Mayer *et al.*, 1991). However, *fackel* seedlings lack the hypocotyl but not the root. In *monopteros* mutants the basic defect affects the central region and the seedlings lack hypocotyl, root and root meristem (Mayer *et al.*, 1991). Taken together these data have been interpreted as indicative of a role of the basal region of the embryo in induction of specification of the adjacent cells as root initials (Jürgens *et al.*, 1995). *raspberry 1* mutants follow a normal embryogenic path until the end of the globular stage period when they fail to undergo the transition to heart stage (Yadegari *et al.*, 1994). The embryos remain globular and the suspensor becomes embryogenic, developing into a mirror image of the apical embryo. Tissue-specific markers which are absent in the suspensor show similar expression pattern in both embryos. In *twin* mutants small morphological alterations of the embryo are accompanied by the formation of a second embryo from proliferating suspensor cells, the two embryos becoming linked by a bridge made of suspensor cells (Vernon & Meinke, 1994). These observations reveal the existence of interactions between the embryo proper and the suspensor (Vernon & Meinke, 1994; Yadegari *et al.*, 1994). *twin2* mutants are characterized by arrest of the apical cell or its derivatives after the first or second division of the zygote (Zhang & Somerville, 1997). The basal cell derivatives initiate the formation of multiple embryos. In the minority of seeds that develop viable embryos a large proportion of these occur as partial or complete duplicates. It has been suggested that the embryo proper exerts an inhibitory influence over a default

embryogenic pathway for the suspensor which is disrupted in these mutants (Vernon & Meinke, 1994; Yadegari *et al.*, 1994; Zhang & Somerville, 1997).

The ability of cells to perceive their position in the developing organism and adapt their gene expression programmes accordingly implies the existence of signalling mechanisms that support cellular interaction. Little is known about these mechanisms. Growth factor-mediated regulation of cell fate has been documented. Liu *et al.* (1993) using inhibitors of polar auxin transport showed its requirement for the establishment of bilateral symmetry in the *B. juncea* embryo. Globular stage embryos treated with the inhibitors developed as phenocopies of the *A. thaliana pin-1* mutants where cotyledon primordia fuse into a single collar-shaped organ. Upon examination, auxin polar transport was found to be disrupted in the *pin-1* mutants, its absence being the cause for the observed shape defect (Liu *et al.*, 1993). Another growth regulator abundantly implicated in control of expression of numerous genes during the maturation/desiccation stages is abscisic acid, ABA (*e.g.* Thomas, 1993).

Evidence for a substantial role of the extracellular matrix components in developmental control is also accumulating (reviewed in Brownlee & Berger, 1995). A 32 kD endochitinase was shown to rescue somatic embryos of a temperature-dependent mutant carrot cell line *ts11* that fail to develop beyond the globular stage when cultivated in embryo-inducing conditions at the non-permissive temperature (de Jong *et al.*, 1992). The endochitinase was a glycosylated protein isolated from an extracellular extract of a non-mutant cell line. The effect of this enzyme could be mimicked by the addition a lipooligosaccharide known to be a nodulation factor from *Rhizobium* (De Jong *et al.*, 1993). This was interpreted as indicative that the endochitinase might be involved in generation of plant homologues of the *nod* factors.

In summary the accumulated evidence indicates that although the clonal history of a cell may contribute to the establishment of its identity, its position in the multicellular embryo plays an important role in the process. The required cell-to-cell communication seems to involve both long-range mechanisms (involving growth control substances) and short-range interactions.

1.3 Current methodology for identification and cloning of genes

The understanding of processes often requires knowledge on the participant elements. This is true of biological processes such as development where the structural and functional characterization of the genes involved and the RNA and proteins they encode is fundamental to their study. Our present ability to study these various elements and investigate their interactions hinges largely on the possibility of identifying and cloning the relevant genes. Both tasks that have been greatly facilitated by an increasing number of molecular genetics strategies and techniques available. Below is an overview of the main techniques presently available to identify and clone genes of interest.

1.3.1 Strategies based on the products of gene expression

An array of strategies and methods has been devised to clone genes of interest that rely on the use of the products of their expression, be it RNA or proteins. The relative abundance of their transcripts, or the ability to specifically isolate them, or the use of the structural properties of their protein products have all been exploited to clone genes. Some of the techniques used are best suited to target specific genes whereas others aim at cloning anonymous genes whose expression is usually associated with a particular population of cells, as discussed below.

1.3.1.1 RNA-based methods

RNA transcripts of genes or their cDNA equivalents can be directly used to screen cDNA or genomic libraries to isolate relevant clones. Alternatively, cDNAs can be prepared by RT-PCR and directly cloned. Subsequent screening of genomic libraries with probes derived from the cDNAs clones allow for the isolation of the corresponding genes. The key element in the choice of this strategy is the ability to isolate the transcripts of interest.

1.3.1.1.1 Direct isolation

The first plant genes to be cloned were distinguished by the high abundance of their transcripts in specific tissues and included those from seed storage proteins, ribulose 1,5-bisphosphate carboxylase, and the leghaemoglobin found in root nodules (*e.g.* Burr & Burr, 1980; Truelsen *et al.*, 1979). The elevated proportion of the transcripts in the total polysomal fraction permitted their specific enrichment and subsequent use for screening libraries or for synthesis and cloning of cDNA. An early example, was the cloning of nuclear sequences encoding zein proteins of the maize endosperm (Burr & Burr, 1980). Taking advantage of the association of zein-encoding polysomes with protein bodies, Burr and Burr (1980) preferentially isolated the corresponding mRNA which was further enriched by size fractionation in sucrose gradients. The highly purified zein mRNAs were then reverse transcribed and used to clone double stranded cDNAs. The very high abundance of a specific class of transcripts encoding leghemoglobin in soybean root nodules was also exploited by Truelsen *et al.* (1979) to clone their cDNAs. One of these cDNAs was later used to isolate clones from a genomic library in a λ vector that contained the respective genes (Jensen *et al.*, 1981).

Clearly this approach is limited to the few genes which have high expression levels in specific cell populations or which otherwise allow for selection of the corresponding RNA or DNA sequences. For example, the distinct buoyant density of the tandemly repeated nuclear rDNA units of the crucifer *Brassica rapa* permitted their specific isolation and subsequent cloning (Rocha, 1991).

1.3.1.1.2 Use of differential expression

Several methodologies have been developed to obtain clones for genes which are differentially expressed in distinct populations of cells. The sample materials may differ in their exposure to environmental or physiological conditions, their tissue or organ origin, their developmental stage, or even in genotype. Commonly, these strategies aim at the isolation of anonymous genes.

a) Differential screening

This is the simplest strategy to obtain cDNA clones for genes which are differentially expressed in two samples. As a first step the method requires the isolation of polysomal RNA from both samples. A cDNA library is then constructed from the RNA sample presumed to contain the sequences of interest. To identify clones of genes differentially expressed, the library is plated at low density and the plaque/colony DNA is transferred to duplicate membranes. These membranes are separately hybridized with cDNA probes prepared from each of the two pools of RNA. The process enables the identification (and subsequent isolation) of those clones corresponding to transcripts present in different relative abundance in the two RNA samples as indicated by the intensity of the hybridization signal.

Because of its simplicity this technique has often been used for cloning differentially expressed genes including several under developmental control. For example, to isolate genes with tissue-specific expression in tomato pistils Gasser *et al.* (1989) prepared cDNA libraries from these organs at three distinct developmental stages. Each library was screened with cDNA probes derived from RNA of seedlings (representing vegetative organs) and with the same RNA used to prepare it. Nine clones were found to hybridize more strongly to the pistil RNAs and were selected for further analysis. All nine were subsequently demonstrated to be under tight spatial and temporal regulation by Northern blot hybridization.

In a more recent application of differential screening to the identification of developmentally regulated genes of *A. thaliana*, Park *et al.* (1998) constructed a cDNA library from senescent leaves. The library was then differentially screened with single-stranded cDNA probes prepared from senescent and from non-senescent leaves enabling the identification of four cDNA clones corresponding to genes activated during senescence.

A significant drawback of this technique is its relatively low degree of sensitivity which poses difficulties or even prevents the detection of genes encoding low-abundance or rare messages. For example, Scott *et al.* (1983) found that in a polysomal RNA population from mammalian cells, those mRNAs representing less than 0.1 % of the total failed to be detected by this technique. An additional difficulty is a requirement for the pattern of expression of the genes to be confirmed by other methods such as Northern blot hybridization.

b) Cloning by subtraction hybridization

A common theme to the plethora of variants of subtraction methods available is the hybridization of a tracer containing the sequences of interest (RNA or cDNA) obtained from a first source of material with an excess of sequences derived from a second population of cells (the driver). This hybridization coupled with a step to eliminate the hybrid molecules and the excess driver enriches for sequences that are unique to or are comparatively up-regulated in source of the tracer. Reiteration of the process further increases the relative proportion of the sequences of interest in the mixture. These sequences can then be directly cloned creating subtracted libraries, and/or used to screen libraries in order to identify and isolate the respective genes. Subtractive hybridization techniques are very sensitive in the identification of genes which are differentially expressed, even when their steady-state levels are low. This sensitivity is enhanced when the subtractive hybridization procedure is coupled with PCR for amplification of cDNAs from rare transcripts.

Subtractive hybridization was first devised as a technique for cloning genes by Scott *et al.* (1983) when faced with the sensitivity difficulties associated with differential screening (see above, al. a). These authors illustrated its use in the identification of transcripts that were produced exclusively or more abundantly in SV40-transformed mammalian cells relative to non-transformed cells. To this end, single-stranded cDNAs were prepared from poly(A)⁺ RNA of both cell populations. The cDNA obtained from non-transformed cells was linked to a solid cellulose matrix and repeatedly hybridized to that derived from the SV40-transformed cells and the subtraction performed. The use of the subtraction tracer as a probe to screen a cDNA library from transformed cells resulted in the identification of relevant clones.

An alternative scheme utilizing the power of PCR amplification was employed in the detection of genes up-regulated during dedifferentiation of *Nicotiana glauca* pith cells in *in vitro* culture when exposed to 2,4-D (Cecchini *et al.*, 1993). In this instance, prior to subtraction, the cDNAs were prepared from differentiated and dedifferentiating pith cells. Linkers were added to the cDNAs that allowed their amplification by PCR with linker-specific primers thus compensating for the low amounts of material available for the subtraction procedure. Transcripts for two of seven major cDNA species present in the subtracted library were found by Northern blotting to have their steady-state levels substantially increased in a period of 4 hours during dedifferentiation.

A distinct use of PCR amplification allied to a subtraction strategy was employed by Phillips (Phillips & Huttly, 1994) to clone genes which are induced by the gibberelin GA₃ in *A. thaliana*. Taking advantage of the existence of the extreme dwarf mutant *gal*, the authors isolated mRNA from two samples of seedlings of the mutant exposed to either GA₃ or to a control application. Both mRNA mixtures were used as drivers for subtraction. Tracers consisted of cDNAs prepared from either of the mixtures by priming with an adapter primer.

After the subtraction procedure, a cDNA library was prepared from the GA₃-treated sample by poly-G tailing of single-stranded cDNA and subsequent PCR amplification followed by cloning. This library was screened with probes prepared from both subtracted cDNAs. Two clones corresponding to transcripts with substantially different steady-state abundances in the treated and untreated plants were isolated.

c) Differential display

In 1992 Liang and Pardee devised a method, differential display, for visualizing differences in the gene expression patterns of two populations of mouse cells and cloning the relevant genes. In differential display, partial random libraries of single-stranded cDNAs are prepared by reverse transcription, usually by priming in the poly(A)⁺ tail. Subsets of the cDNAs are partially amplified by PCR with a poly(T)-tail primer in combination with anonymous primers. Libraries from different populations of cells are amplified in simultaneous and the products electrophoretically resolved in parallel. This enables the detection of bands of fragments unique to one or a subset of the libraries used indicating their likely correspondence to transcripts absent in the other populations of cells. Because they function as tags for genes, the fragments in the bands can then be isolated and cloned or used to screen cDNA or genomic libraries. Alternatively, the isolated fragments can be directly sequenced. An advantage of this technique, in addition to its sensitivity, is its enabling of detection of differences of gene expression in all populations under analysis.

The first application of differential display to cloning of plant gene sequences was reported by Wilkinson *et al.* (1995). The technique was used to visualize differences between two mRNA populations isolated from unripened and ripe strawberry fruits. Fragments of five bands unique to ripening plants were cloned and used to isolate corresponding cDNAs later shown to derive from distinct genes. Northern blot analysis showed that the expression of three of the genes was restricted to the maturing fruits.

Differential display has been used to clone other developmentally controlled genes including some expressed during embryogenesis. Using this technique with RNA samples extracted from young *B. napus* embryos (transition and heart stages), from older embryos, shoot apex, and mature leaves it was possible to identify one band of amplified 500 bp fragments unique to the young embryos (Heck *et al.*, 1995). A corresponding cDNA and a related one, as well as genomic clones, were isolated that encoded members of the MADS domain family of regulatory elements whose mRNA accumulated primarily in embryos.

1.3.1.2 Cloning genes from their protein products

When available, the purified proteins have been used in two main ways for cloning of their genes. One exploits the specific interaction of the proteins with antibodies previously raised against them or against epitopes derived from them. This strategy involves the screening of a

protein expression library with specific antibodies. The interaction of the antibodies with the cognate protein permits the detection of cDNA clones encoding the latter.

Numerous genes have been isolated by resorting to antibody screening, including some likely to have regulatory roles in gene expression. For example, to isolate *Arabidopsis* cDNA clones for proteins binding to the G-box, a well characterized *cis*-regulatory element found in numerous plant promoters, Lu *et al.* (1992) prepared a monoclonal antibody against an unknown G-box binding activity and screened an expression library with this antibody. This strategy allowed the isolation of a cDNA clone for the activity. It encoded a protein with similarity to protein kinase C inhibitors and to activators of tyrosine and tryptophane hydroxylases from mammals.

The use of antibodies against a well-defined epitope for cloning purposes is illustrated by the work Lange *et al.* (1994). Based on the partial microsequencing of purified gibberellin 20-oxidase these authors raised antibodies against a specific nine-residue peptide. Screening of a *Cucurbita maxima* expression library identified one cDNA clone which was found to encode the enzyme. A similar approach was followed by Kato *et al.* (1996) who screened a λ gt11 expression library of pumpkin for clones encoding 3-ketoacyl-CoA thiolase. Screening was performed with antibodies raised against a 28-residues peptide previously defined by microsequencing.

Partial microsequencing of small amounts of purified protein lies at the base of a different set of strategies to clone the corresponding genes. From the partial peptide sequences degenerate mixes of oligonucleotides are designed that can then be used to screen cDNA or genomic libraries. Because screening with degenerate oligonucleotides is technically challenging, a more popular technique is the amplification of cDNA or genomic segments by PCR primed with the oligonucleotide mixes. The amplification of the segments can be performed directly from DNA or it may follow prior reverse transcription of the relevant mRNA with an appropriate primer (RT-PCR). Libraries are subsequently screened with the amplified fragments for isolating cDNA or gene clones.

The utilization of oligonucleotide-based strategies is exemplified by the cloning of the tomato ACC synthase and the expansins of cucumber: Van Straeten *et al.* (1992) used partial knowledge of the tomato ACC synthase sequence obtained by microsequencing of several tryptic peptides of the purified enzyme to design degenerate oligonucleotides. Screening of a tomato cDNA library with these oligonucleotides resulted in the recovery of two clones that proved to encode the ACC synthase enzyme. To isolate cDNA clones for the S1 and S2 expansins of cucumber Shcheban *et al.* (1995) screened an appropriate library with two fragments obtained by RT-PCR. The oligonucleotides used for the RT-PCR amplification were designed from partial peptide sequences derived from purified expansins.

The isolation of genes of interest based on their protein products is limited by the ability to isolate and purify the latter, and do so in sufficient quantity and quality to allow for antibody elicitation and/or microsequencing.

1.3.2 Homology-based cloning

The cloning of many genes has relied on their structural similarity to other previously cloned homologues from the same or distinct species. This has been accomplished by screening genomic or cDNA libraries with probes derived from the cloned genes or from their cDNAs. Alternatively, conserved regions of the genes can be identified and exploited for the identification of the sought clones following one of two strategies based on the use of oligonucleotides. Also, screening with antibodies raised against the protein product of the homologous gene can be used to isolate cDNAs of interest from expression libraries. For those species for which collections of ESTs are available, cDNAs of interest may be retrieved following simple comparative sequence analysis.

The foundation of the homology cloning methods also defines their limitations: these methodologies can only be applied to genes for which one or more previously cloned homologues or protein-specific antibodies are available. Furthermore, the degree of divergence between the target gene and the homologue must be sufficiently reduced to allow for detection.

1.3.2.1 Screening with homologous genes and cDNAs

The isolation of a gene of interest can be greatly facilitated when clones for homologous genes are available. Provided that their degree of divergence is sufficiently small, as can be tested by Southern blot hybridization, the homologous genes or cDNAs can be directly used for screening libraries. This type of approach is illustrated by the isolation of three cDNAs of *B. oleracea* with structural similarity to *flo*, an homeotic gene of *Antirrhinum majus* expressed early in floral initiation (Anthony *et al.*, 1993). In this case, a probe prepared from a genomic clone of *flo* was used for screening of a *B. oleracea* cDNA library. Library screening with an heterologous cDNA probe was the strategy successfully used by Liang *et al.* (1991) to isolate clones representing ACC synthase genes of *A. thaliana*. The existence of four genes encoding this key enzyme in the production of the hormone ethylene was recognized in this manner.

Different genes are subjected to distinct evolutionary pressures hence diverging at distinct rates. The higher the degree of conservation of genes, the more phylogenetically distant can be the source of the gene or cDNA probes used for heterologous screening. For instances, Learned *et al.* (1989) used a 0.3 kb fragment of the cDNA encoding yeast 3-hydroxy-3-methylglutaryl coenzyme A reductase highly similar to its hamster homologue to isolate a cDNA clone encoding the corresponding enzyme of *A. thaliana*.

1.3.2.2 Two uses for oligonucleotides

For faster evolving genes two useful strategies have been employed which take advantage of the unequal rate of change experienced in different regions. Both approaches rely on the identification of well conserved segments of proteins by comparative analysis of sequences from various sources. Mixes of degenerate oligonucleotides are then designed from the conserved segments. The two strategies differ in the use of the oligonucleotide mixes. They are identical in their application to the methods of cloning genes from sequences of their encoded proteins described in Section 1.3.1.2. In one type of application, cDNA or genomic libraries are directly screened with labelled oligonucleotides for direct isolation of clones that potentially encode the proteins of interest. This strategy was first employed in the isolation of a cDNA for a *ras*-related gene, *ral*, from a simian library (Chardin & Tavitrán, 1986). In this instance the oligonucleotides used were derived from a seven-residues peptide sequence strictly conserved in all *ras* proteins from yeast to man.

An alternative strategy consists in the utilization of two sets of degenerate oligonucleotides derived from two distinct conserved segments of the protein to prime PCR amplification of the equivalent segment in the gene of interest. The amplification can be directly initiated from genomic DNA or cDNA, or following a reverse transcription step (RT-PCR) of RNA. The amplified fragments are then used in the screening for cDNA or genomic clones. For cloning an *Arabidopsis* functional homolog of the p34^{cdc2} protein kinase Ferreira *et al.* (1991) first compared all known sequences of cdc2 proteins and identified the conserved regions. From each of two of the conserved segments a set of degenerate primers were designed. The two sets were used to prime the amplification of a fragment of a cDNA that was shown to be similar to p34^{cdc2} protein kinases. Full-length cDNAs and a corresponding gene were isolated from libraries after screening with the fragment.

1.3.2.3 Antibody screening

This approach offers two advantages relative to the other methods. It can exploit evolutionarily more distantly related genes given that the rate of change of proteins is lower than their encoding sequences. In addition, the availability of antibodies to an heterologous protein product may make redundant the existence of clones for the cognate genes. The practical application of this concept is essentially identical to cloning by screening expression libraries with antibodies raised against the cognate product of the gene (Section 1.3.1.2). The only distinction lies in the origin of the epitopes against which the antibody was raised.

An early example of the use of this approach to clone sequences from a plant gene was the successful isolation of cDNA clones for an *A. thaliana* polyubiquitin gene following screening of a λ gt11 expression library with anti-polyubiquitin antibodies previously prepared against oat and human ubiquitin (Burke *et al.*, 1988).

1.3.2.4 Homology-based selection of ESTs

For a few species including the model plant species *A. thaliana* and *Oryza sativa* growing collections of ESTs are publicly available (Meinke *et al.*, 1998; Newman *et al.*, 1994).

Comparative search of the EST sequences with sequences derived from genes for which homologues are being sought allows for the identification of likely cDNA candidates. The use of comparative analysis programs of the BLAST family (Altschul *et al.*, 1990) allows the search to be conducted at the nucleotide sequence level or at the more sensitive protein level following automatic translation of the sequences. The selected cDNA candidates can then be retrieved from the depositories of the EST collections and their sequence analyzed further to establish reliably their homology to the heterologous gene.

1.3.3 Genetic/Functional strategies

1.3.3.1 Complementation cloning

The evolutionary relatedness of organisms can be exploited to clone a gene whose function and/or phenotype is well defined and shared between two species. The gene of one species can be identified by the expression of a wildtype, tagged allele in an heterologous species carrying a recessive or co-dominant mutant of its homologous gene. This genetic complementation strategy usually involves the construction of an expression library in a bacteria or eukaryote host, normally *E. coli* and yeast, coupled with a screening procedure for detection of transformants where the complementation occurs.

An example of the application of this technique to the isolation of cDNAs for plant transcripts was the identification of cognate cDNA clones for superoxide dismutase of *A. thaliana* and *N. plumbaginifolia* in libraries constructed in a superoxide dismutase-deficient strain of *E. coli* (Van Camp *et al.*, 1990). Recently, Umeda *et al.* (1998) cloned an *A. thaliana* cyclin-dependant kinase-activating kinase (CAK) by genetic complementation in yeast. In this particular instance an yeast strain carrying a CAK temperature-sensitive mutation was used to prepare an expression library of cDNA clones of *A. thaliana*. By transferring the library to the non-permissive temperature the complemented transformants were identified.

1.3.3.2 Dihybrid system for detection of interacting proteins

Fields and Song (1989) provided a conceptual jump in cloning strategies when they explored the functional properties of the yeast GAL4 transcription factor to detect protein-protein interactions. The system is based on the expression in a suitable yeast strain of two fusion proteins. One is composed of the N-terminal DNA binding domain of GAL4 fused with a specific, known protein. The other protein is a fusion between the transactivating C-domain of GAL4 and the protein whose interaction with the known protein is to be tested. Interaction

between the two fusion proteins is detected by its mediation of the expression of a positive selection gene requiring GAL4 transactivation at its promoter. This and similar systems can be used for the isolation of genes encoding proteins that interact with a specific protein by construction and screening of an expression library of fusions of the proteins to be tested.

Recently two-hybrid systems have been employed to obtain cDNA clones for a variety of interacting proteins. For instances one such system was used to isolate cDNA clones of *A. thaliana* encoding proteins interacting with GTP-binding Ran proteins, thought to be necessary for protein import into the nucleus and for the onset of mitosis (Haizel *et al.*, 1997). Two cDNAs encoding distinct proteins, At-RanBP1a and At-RanBP1b, were recovered in this manner.

1.3.3.3 Monohybrid system

Wang and Reed (1993) developed a system for detection of proteins interacting with DNA elements based on the transactivating properties of the yeast GAL4 transcription factor. In this yeast monohybrid system a library of protein fusions to the GAL4 transactivating domain is made. The library is then transformed into an yeast strain which carries a selection marker (*his3⁻*) and a normal but inactive allele (*HIS3*) in a plasmid. Expression of the normal allele requires binding of a fusion GAL4 to the promoter region of *HIS3* containing the selecting DNA elements. This binding can only be mediated by proteins that specifically interact with the test DNA elements. Therefore, by growth of the yeast library in the appropriate media it is possible to isolate those fusion cDNA clones of interest. These can in turn be used to isolate cDNA and genomic clones corresponding to the unknown protein.

A cDNA clone encoding a transcription factor of *Arabidopsis*, ARF1, that binds auxin response elements was identified and isolated with the aid of a monohybrid system (Ulmasov *et al.*, 1997). For the isolation, repeats of the previously defined minimal auxin response element TGTCTC were used in the promoter region of the selection gene.

A prior detailed characterization of promoter elements is not required for the identification of interacting proteins. For example, a monohybrid system was also used to identify nine clones corresponding to transcripts encoding two leucine-zipper type proteins that interact with the proximal promoter region of *Dc3* (Kim *et al.*, 1997). This gene is a member of the *lea* class abundantly expressed in somatic and zygotic embryogenesis, normally embryo-specific but also inducible with ABA.

1.3.3.4 Phage display and related techniques

The ability of fusion proteins to incorporate into bacteriophage coats without disrupting their infectivity has been explored for detecting and selecting those that interact with a specific target, as pioneered by Smith (1985). Fusion proteins can be made to the coat proteins of bacteriophages, including to the N-terminus of the major coat protein pVIII (present in ~2800

copies) or to the tail protein pIII (5 copies) of the filamentous phage (Smith, 1985). By constructing a library of fusion proteins and submitting it to cycles of selection and amplification in *E. coli*, it is possible to enrich substantially for those fusions interacting with a given target. The system was first used to select and enrich for a specific fusion protein by antibody selection (Smith, 1985). To date phage display has not been used for cloning genes of plant origin.

A related strategy consists of the functional screening of expression libraries. An interesting application of this concept was provided by the screening of an *Arabidopsis* expression library in λ ZAPII for proteins that mediate the interaction of the 26S proteasome with proteins tagged for degradation by multiubiquitination (van Nocker *et al.*, 1996). Screening was performed with labelled multiubiquitin chains, resulting in the isolation of a clone encoding MBP1, a protein homologous to subunit 5a of the human 26S proteasome.

1.3.4 Methods based on chromosomal differences

1.3.4.1 Cloning of deleted genes

When null mutations resulting from deletion of the gene of interest are available it is often possible to clone the corresponding gene by genomic subtraction. The method was developed by Straus and Ausubel (1990) who used it to identify clones corresponding to a 5 kb deletion in a yeast mutant. It essentially consists of subtracting the common sequences of the wildtype genome (tracer) of the corresponding sequences in the deletion mutant genome (driver). The subtraction procedure can be reiterated to ensure that only DNA corresponding to the deletion in the mutant is recovered. Following the addition of adapters to its termini the subtracted DNA is amplified by PCR and then used to probe a wildtype genomic library.

An important limitation of the method is the requirement for viable homozygous deletion mutants. Also, the characteristics of the genome (complexity and sequence redundancy) and the size of the deletion are important factors affecting the usage of the technique.

1.3.4.2 Representational difference analysis (RDA)

This is a very powerful and sensitive method to clone segments differing in size between two otherwise largely identical genomes capable of detecting the exclusive presence of a single integrated virus copy in one of the targets. Developed by Lisitnyn *et al.* (1993), it is carried out in three steps. First, a fractional amplification of the genomes is performed with a whole genome PCR. This creates representations of the genome. The second step consists in a subtraction hybridization of one of the genomes with a great excess of the representations from the other genome. Finally, adaptors are added to the termini of the subtracted fragments which

are then amplified by PCR following filling, with primers designed for the linkers. The amplified fragments can subsequently be cloned and analyzed.

In plants RDA was first applied to identify male-specific restriction fragments in the dioecious plant *Silene latifolia* (Donnison *et al.*, 1996). The corresponding fragments were then analyzed in mutant plants characterized by the presence of hermaphroditic or asexual flowers. This permitted establishing correlations between cytological abnormalities and the absence of specific fragments in the white campion mutants allowing their physical mapping in the Y chromosome near a suppressor gene of carpel formation.

1.3.4.3 Cloning genes from their physical location

a) Positional cloning

Frequently the only information on a gene of interest is merely genetic relying on a scorable phenotype with no known clones of homologues from other species being available. Cloning of such genes is feasible if they can be located on a genetic linkage map (reviewed in Tsui & Estivill, 1991). This process which has been variously designated as positional cloning, reverse genetics and forward genetics, involves three distinct steps. The starting point for positional cloning is the establishment of the location of the trait of interest in a genetic map and the identification of markers tightly linked to and flanking it. The next step consists of isolating the physical region thought to contain the trait. By analyzing this region in two different alleles (*e.g.* a mutant and a wildtype) it is possible to determine where the difference is located, consequently identifying the gene. Complementation analysis is used to confirm the identity of the gene, where possible.

Several difficulties are associated with positional cloning which are directly related with the complexity and organization of the genome. For organisms with complex genomes the gaps between the gene of interest and nearby markers can be substantial. As a consequence, genomic clones containing the nearby markers may not be sufficiently large to also include the region of the trait. Isolation of this region will then require chromosomal walking whereby clones of overlapping genomic segments are sequentially isolated by screening libraries with probes from the terminal regions of a previously mapped clone. This process is reiterated until the segment containing the trait of interest is known to be included in one of the clones. Besides being time-consuming, chromosomal walking is often hindered, even prevented, by the presence in many genomes of substantial proportions of interspersed repeated sequences.

b) Molecular markers linkage and high density mapping

The map-based cloning of a gene is facilitated by its closer linkage to a marker. The scarcity of markers in classical genetic maps makes them generally insufficient for this procedure. Denser

linkage maps can be more easily obtained with molecular markers. Furthermore, scoring of molecular markers is straightforward and is not affected by complexities resulting from genotypic relationships or epigenetic phenomena that may affect linkage analysis of classical genetic markers. For these reasons, the construction of maps for molecular markers has received significant attention in recent years.

The first maps incorporating molecular markers were based on RFLPs. Since the initial suggestion of use of RFLPs as co-dominant markers in linkage studies by Botstein *et al.* (1980), a panoply of other types of molecular markers useful for mapping have been made available. Included are the (generally) dominant Random Amplified Polymorphic DNA (RAPD) markers, first applied to the construction of linkage maps of the human and soybean genomes (Williams *et al.*, 1990). Also included are the co-dominant Amplified Fragment Length Polymorphisms (AFLPs) (Vos *et al.*, 1995), and the CAPSs markers developed by Konieczny and Ausubel (1993). The use of SSLPs for linkage studies was first devised from analysis of simple sequence repeats in the *Arabidopsis* genome by Bell and Ecker (1994).

c) Chromosomal landing and synteny-based mapping

In recent years the existence of methodologies for creating high density molecular linkage maps has been exploited to avoid the considerable difficulties often found in chromosomal walking (see above) for map-based cloning of genes. Chromosomal walking can be avoided if genetic markers can be found sufficiently near the gene of interest so that both may be within the span of a single genomic clone (Tanksley *et al.*, 1995). To achieve this, an initial mapping of the trait with existing markers is followed by a new cycle of mapping performed with new molecular markers to find close linkages. For this approach, known as chromosomal landing, new markers in the region of interest can be found by resorting, for example, to Bulk Segregant Analysis (BSA) or Recombinant Inbred Lines (RILs) (Lister & Dean, 1993; Michelmore *et al.*, 1991).

The accumulation of data on the genomes of a variety of species has uncovered significant degrees of conservation of the organization of their genes, particularly among closely related species, referred to as synteny (Devos & Gale, 1997). For example, the gene organization in eleven regions of the *Arabidopsis* genome (corresponding to about 25% of its total) are conserved in *B. oleracea* (Kowalski *et al.*, 1994). This organizational similarity can be exploited to clone genes mapping to the conserved regions.

d) Applications in plants

Unfortunately the genomes of plants are often very large, ranging in size from about 100 Mb in *Arabidopsis* to 12,700 Mb (Bennett & Smith, 1976). Furthermore, one of their characteristics is a high level of interspersed repeated sequences (cf. Section 1.2.2). For this reason cloning of plant genes by positional cloning has largely been limited to a few species, in particular the dicot *A. thaliana* and the monocot *Oryza sativa* where a relatively small genome size is

combined with an unusually low content of interspersed repeated DNA content (see Section 1.2.2). In addition, for *Arabidopsis*, relatively dense linkage maps are available. All these attributes make *A. thaliana* a good genetic system for positional cloning. Several *Arabidopsis* genes identified solely on the basis of mutant phenotypes have been cloned by this procedure. Among the first to be cloned were the abscisic acid insensitive locus *abi3* (Giraudat *et al.*, 1992). Genes with important functions in pattern formation in the developing embryo of *Arabidopsis* have also been isolated by positional cloning. These include *GNOM* and *KNOLLE* (see Section 1.2.4.2) which were found to encode, respectively, a protein featuring a domain with similarity to the yeast secretory protein Sec7p, and a protein with similarities to syntaxins, a family of proteins involved in vesicular trafficking (Busch *et al.*, 1996). Another gene affecting pattern formation, *MONOPTEROS* (see Section 1.2.4.2), was also isolated by positional cloning (Hardtke & Berleth, 1998). The protein product of the gene includes a nuclear localization signal and a DNA-binding domain.

1.3.5 Gene tagging by insertional mutagenesis

1.3.5.1 Introduction

The recovery of genes of interest can also be accomplished by their modification with a molecular tag that allows their specific identification and isolation. The molecular tags are DNA elements that function by intercalation in the target genome, thus being mutagenic.

Undoubtedly, the most important application of tagging is the creation of new mutants affected in functions of interest and the straightforward recovery of the affected genes. The concept of gene tagging as a cloning tool was developed by Bingham *et al.* (1991) who first recognized that the mobile nature of transposons makes them suitable molecular tags. They first exploited the association of a *copia* element with the *white* locus in a known mutant of *Drosophila melanogaster* to recover sequences from it.

The power of gene tagging as a cloning method resides on the ability to create the necessary association between a sequence element, the tag, and the gene of interest, and on the use of the known structure of the former to isolate the latter. This approach targets specific genes identified by an associated trait caused by insertion of the tag. Recovery of the affected gene is straightforward, far more efficient and faster than other targeted strategies, such as positional cloning. Isolation of the target gene involves that of native sequences flanking the tag by a variety of techniques including inverse PCR (IPCR; Triglia *et al.*, 1988), construction and screening of a genomic library of the mutant, or plasmid rescue when the tag includes the required replicative functions (*e.g.* Feldmann and Marks, 1987). Usually, the mutants generated by insertional mutagenesis are of the loss-of-function type and their phenotype is recessive. It is

then possible to identify the affected gene in lethal-mutations in viable hemizygous individuals, a significant advantage of the method.

The advantages of insertional mutagenesis have been multiplied by designing tags that not only can potentially disrupt genes by insertion but also report on the activity of regulatory elements of transcription (*e.g.* Topping *et al.*, 1991; see below, Section 1.3.5.4) and even create gain-of-function phenotypes (*e.g.* Walden *et al.*, 1994; see below, Section 1.3.5.5).

In plants, two types of elements with the appropriate characteristics to function as tags have been used, namely transposons and T-DNA. Their use in the various applications of insertional mutagenesis is the subject of the sections below.

1.3.5.2 Transposons as molecular tags in plants

Transposons were first identified by Barbara McKlinton in maize and since numerous examples of these elements have been identified in higher plant species (reviewed in Doring & Starlinger, 1986). Characteristically, transposons are DNA elements with short inverted repeats at their termini which are necessary for transposition. Their integration is accompanied by duplication of a short stretch of the target site. Excision of the transposon often does not restore the site of integration. Transposition requires the activity of enzymes, transposases, encoded by some transposable elements. These fully functional transposons are designated autonomous elements. Other transposable elements, so-called non-autonomous elements, do not encode a functional transposase. For this reason they are dependent for transposition on the enzyme activity supplied by a related autonomous element.

Several characteristics of transposons are useful in gene tagging strategies including, for some, the ability to transpose with relative high frequency and creating large numbers of different mutants, and the potential reversion of mutation by excision. The latter is often sufficient as proof that the mutation considered was caused by the insertion of the element.

The best studied transposons in plants include the autonomous/non-autonomous element systems *Ac/Ds*, *Spm/dSpm* (*En/I*) and *Mutator* of maize, and the autonomous elements *Tam1* and *Tam3* of *Antirrhinum majus* (Walbot, 1992). All of these transposons have been used to clone genes of their native plants. The first application of transposon tagging to clone a plant gene was reported for the *Bronze* locus of maize (Fedoroff *et al.*, 1984). In this instance a *bronze* mutant of maize previously shown to be caused by an insertion of an *Ac* transposon, was used to isolate segments flanking the insertion element from a genomic library. These segments allowed the isolation of the corresponding wildtype region containing the *bronze* locus.

Several transposons have been shown to be active in heterologous species, as first shown for *Ac* in tobacco (Baker *et al.*, 1986). This has permitted gene tagging with heterologous transposon systems, a practice that offers several advantages. Among these is the absence of other copies of the transposon in the genome which can interfere with the recovery of the tagged

genes (Walbot, 1992). In addition, the activity of the transposase can be regulated and the element modified.

The *Ac/Ds* system has been the most widely utilized for heterologous transposon tagging, particularly in *A. thaliana* (van Sluys *et al.*, 1987). For this reason the system and its usage are reviewed in more detail here. The autonomous *Activator* element is 4565 bp and encodes a transposase 807 residues in length and requires a 200 bp segment of the terminal repeats for mobility (Bancroft *et al.*, 1992). It has also been introduced in tobacco, tomato, carrot, potato, petunia and rice (Baker *et al.*, 1986; Chuck *et al.*, 1993; van Sluys *et al.*, 1987; Walbot, 1992) where it actively transposes. In *Arabidopsis*, although active, the level of transposition is low and insufficient for productive gene tagging strategies (van Sluys *et al.*, 1987). This difficulty can be overcome by modifications of the 5' UTR of the transposase transcript (Bhatt *et al.*, 1996), or by increasing its expression, for example by placing it under the control of a strong promoter (Grevelding *et al.*, 1992). The use of two-component systems where a stable, non-transposable *Ac* element or another transposase source, and a *Ds* element are simultaneously present in the genome, have been very useful in this respect (Bancroft *et al.*, 1992; Grevelding *et al.*, 1992; Long *et al.*, 1993). Two important features of the two-component systems are the ability to control the level of expression of the transposase and therefore of transposition, and the possibility to stabilize integration events of interest by segregation of the *Ds* and the *Ac* elements.

A long recognized feature of *Ac/Ds* transposition in maize, which has also been observed in tomato and *Arabidopsis*, is a relatively high frequency of insertion of the transposon at sites linked to the donor sites (Bhatt *et al.*, 1996; Osborne *et al.*, 1991; Walbot, 1992). Consequently, the *Ac/Ds* system is suitable not only for non-targeted insertion and mutagenesis, but also for targeted insertion whereby a target gene is mutagenized by transposition from a nearby site. To take advantage of this property, Long *et al.* created recently a population of *Arabidopsis* with *Ds* insertions in all five chromosomes and confirmed the usefulness of the system for targeted transposition (Long *et al.*, 1997).

Several genes of *A. thaliana* have been cloned following transposon element-mediated tagging. For instances, two embryo-lethal mutants generated with two distinct two-component systems were isolated which aborted at the globular stage to heart stage transition. The mutants harboured the inserted *Ds* elements in the nuclear genes for two plastid proteins, namely a homolog of the ribosomal S16 protein (Tsugeki *et al.*, 1996), and a glycyl-tRNA synthetase (Uwer *et al.*, 1998).

1.3.5.3 T-DNA tagging

The transfer DNA (T-DNA) portion of the tumour- and hairy root-inducing plasmids (Ti and Ri) of *Agrobacterium* species is another naturally occurring insertional mutagen (reviewed in Zambryski *et al.*, 1989). In infected plant cells, T-DNA is transferred from the bacteria to the

nucleus where it integrates into the genome. These processes are not dependent on T-DNA-encoded functions, although they require the presence of at least one of the 24 bp imperfect terminal repeats found at its left and right borders, the transfer taking place in a polar manner starting at the right border. Therefore, engineering of T-DNA has been possible including the removal of unwanted T-DNA encoded functions and the incorporation of other desired DNA sequence elements, such as selectable markers, making them suitable for a variety of purposes including gene tagging.

T-DNA-mediated mutagenesis has been performed in several species among them *N. tabacum*, *N. plumbaginifolia*, tomato, potato, and, particularly, *Arabidopsis* (e.g. André *et al.*, 1986; Feldmann, 1991; Feldmann & Marks, 1987; Koncz *et al.*, 1989; Lindsey *et al.*, 1993). In *Arabidopsis*, numerous mutants have been generated either by seed transformation (Feldmann, 1991; Feldmann & Marks, 1987), co-cultivation in tissue culture (e.g. Lindsey *et al.*, 1993), or *in planta* transformation (e.g. Clough & Bent, 1998). A wide range of altered phenotypes has been observed in the transformed plants. In a collection of 13,000 transformants, scoring for obvious visible phenotypes revealed the presence of several classes of mutants characterized by seedling-lethality, size variation, defective embryos, reduced fertility, and with pigment, flower, root or trichome alterations, dramatic shape variations, and physiological mutants (Feldmann, 1991; Forsthoefel *et al.*, 1992). Unlike transposon-induced mutations, the lack of mobility of T-DNA requires that the demonstration of direct linkage of a phenotype to an integrated T-DNA be made by means other than reversion. Thus, co-segregation of the mutant phenotype with the T-DNA (normally followed with reference to a selection marker) is usually investigated. However, final proof of a causal link between a T-DNA tagged candidate gene and an affected function requires functional complementation of the defect with its wildtype allele. Indeed, many of the mutations detected following transformation have been shown not to be linked to and directly caused by T-DNA insertion (Feldmann, 1991; Forsthoefel *et al.*, 1992).

The wide mutational spectrum obtained with this type of mutagenesis indicating randomness in integration suggests that at least a large part of the genome is susceptible to insertional modification by T-DNA (Feldmann, 1991; Forsthoefel *et al.*, 1992; Lindsey *et al.*, 1993; Topping *et al.*, 1991). However, in a high proportion of transformants obtained with T-DNAs designed to detect transcriptional and translational fusions (see below, Section 1.3.5.4) it has been shown that the integrations are localized in transcribed regions of the genome, irrespective of its size. This has been interpreted as indicative of a preference for integration at transcribed or potentially transcribed sequences.

The simplicity of recovering T-DNA tagged segments of the genome has resulted in the cloning of several genes including some involved in developmental processes in *Arabidopsis*. Among the first to be cloned following T-DNA tagging was *GLABROUS1* which participates in trichome development (Marks & Feldmann, 1989). From the same collection of transgenic plants a mutant defective in another gene also required for proper formation of trichomes,

ZWICHEL, was identified. The wildtype gene was isolated and found to encode a member of a family of microtubule motor proteins, the kinesins (Oppenheimer *et al.*, 1997). The mutants *twin2* and *emb30*, an allele of *gnom*, (see Section 1.2.4.2) identified in the same collection of transgenic plants were used to clone the respective wildtype genes (Shevell *et al.*, 1994; Zhang & Somerville, 1997).

1.3.5.4 Promoter and enhancer trapping

The insertion of a molecular tag in the genome offers new opportunities for the identification and isolation of regulatory elements of gene expression. The mutagens can be engineered to potentiate the creation of transcriptional and translational fusions. In these instances, a promoterless reporter or marker gene is placed near the border of the insertion element so that its expression will be dependent on transcription initiated from a native promoter. If the reporter or marker gene lacks an initiation codon, detection of its encoded activity requires the formation of a translational fusion in the appropriate reading frame. In both situations detection of reporter gene activity is expected to be controlled by, and therefore reflect, the activity of genomic regulatory elements of expression, likely native promoters. This is referred to as promoter-trapping, and the mutagens have been referred to as interposons by Topping *et al.* (1991). In a variation of this scheme, potential 3' splice sites are placed upstream from the reporter gene which may lack an initiation codon. This allows for the detection of the reporter activity when the tagging element is inserted within the intron of a native gene, a strategy often designated exon- or gene-trapping (Gossler *et al.*, 1989).

In the first reported use of a promoter trap in plants, an antibiotic resistance gene, the aminoglycoside phosphotransferase II of Tn5, linked to a transcription terminator of opine synthase, was placed in a T-DNA adjacent to its right border repeat (André *et al.*, 1986). Diploid protoplasts of *N. plumbaginifolia* were transformed with the T-DNA by co-cultivation with *Agrobacterium*. Transformed calli were selected that expressed the resistance gene indicative of its activation by non-T-DNA promoters. In similar experiments but with a different resistance marker, the neomycin phosphotransferase gene *nptII*, Teeri *et al.* (1986) showed by Northern blot analysis tissue specificity of expression of the fusion transcripts in *N. tabaccum* plants recovered from resistant calli. Inversion of the orientation of the reporter gene resulting in the abolition of the T-DNA transformation to confer resistance to neomycin was used by Herman *et al.* (1990) to demonstrate that the expression of a *nptII* reporter gene was dependent on regulatory elements of the native genome.

The introduction of reporter genes that encode histologically detectable activities such as the *E. coli* β -glucuronidase (*gusA*), luciferase enzymes, and, more recently, the green fluorescent protein has significantly enhanced the usefulness of promoter trapping strategies. By resorting to these reporters it is possible to follow in time and *in situ* the activities of regulatory elements of expression in different tissues, organs and cells. The *gusA* gene has been the most

widely used reporter because among other properties, it is easy to assay in a very highly sensitive way by fluorimetry allowing quantitative measurements, in addition to its simple histochemical *in situ* detection (Jefferson *et al.*, 1987). In the first combined uses of GUS *in situ* histochemical detection and T-DNA insertional mutagenesis to analyze patterns of reporter gene expression driven by regulatory sequences of the plant genome, Topping *et al.* (1991) found a variety of GUS activity patterns in a collection of tobacco transformants, and similar observation were made in *Arabidopsis* transformants by Kertbundit *et al.* (1991). Reporter activity was observed to have distinct organ and tissue specificity and variable levels in different plants. Similar findings were obtained with another *gusA*-based promoter trap by Lindsey *et al.* (1993) in transgenic *Arabidopsis*, tobacco and potato plants.

A noticeable and generally consistent feature of promoter-trapping experiments in plants is the elevated proportion of transformants where fusion transcripts are detected. Koncz *et al.* (1989) using a T-DNA vector with the aminoglycoside phosphotransferase II (*aph(3')*II) gene as a reporter next to its right border repeat transformed *N. tabacum*, *N. plumbaginifolia* and *A. thaliana*. In all three species about 30% of transcriptional fusions were detected in spite of the significant differences in the sizes of their genomes and content of repeated, non-transcribed, DNA. Herman *et al.* (1990) studied the frequency of activation of a promoterless reporter gene *nptII* in tobacco plants transgenic for a T-DNA-based promoter trap. Over 25% of the plants were resistant to neomycin. Similarly, Kertbundit *et al.* (1991) found fusion transcripts in 54% of a population of *Arabidopsis* transformed with a *gusA*-based promoter trap. In a distinct population 1.6% of transgenics evidenced translational fusions. In a more comprehensive study involving transformants of tobacco, *Arabidopsis* and potato, Lindsey *et al.* (1993) showed that the incidence of GUS activity was high for all three species, albeit variable in different organs. The maximum frequencies of GUS activity detected in a given organ system ranged from 25% in stems of potato (12/48) and 30% in roots of *Arabidopsis* (28/94) up to 92% in flowers of 24 tobacco transformants.

These observations, taking in consideration the average number of T-DNA loci (1 to 3) present in the transformants, have been interpreted as resulting from a bias of T-DNA insertion events to occur transcriptionally active or competent regions of the genome (Herman *et al.*, 1990; Kertbundit *et al.*, 1991; Koncz *et al.*, 1989; Lindsey *et al.*, 1993).

Interposons based on transposable elements have also been utilized. Springer *et al.* (1995) first employed a modified Dc element including a promoterless *gusA* as a gene-trap, in a dual-component system. This same type of system has another significant application in developmental biology, namely in clonal analysis (*e.g.* Scheres *et al.*, 1994); cf. Section 1.2.4.4).

The interposon-mediated identification of regulatory elements of gene expression in a target genome was pioneered in *Drosophila* by transformation with a P-element where the reporter gene, *lacZ*, was linked to a weak promoter (O'Kane & Gehring, 1987). The presence of

the weak promoter enables regulatory elements active at a distance, enhancers, to induce transcription of the reporter. This experimental design is accordingly known as enhancer-trapping. Topping *et al.* (1991) applied this experimental conception to plants by means of a T-DNA interposon where the *gusA* gene was fused to the weak (-90bp)CaMV35S promoter. Of 184 tobacco plants transformed with the enhancer-trap 73% displayed GUS activity which differed in level over a 300-fold range.

The combination of promoter and enhancer-trapping with histogenically detectable reporter activities constitutes a powerful tool in the developmental biologist's set. It enables the selection of genes and regulatory elements for isolation and analysis according to their patterns of activity, irrespective of the possible absence of a visible phenotype. Furthermore, these patterns of expression can yield clues as to the function of the tagged gene, a feature particularly useful when a recessive mutation is lethal.

The external regulation of their activity, in a cell-autonomous manner, confers on the reporters the attributes of functional markers. This is significant for and has direct application in the investigation of the regulatory circuitries affecting their expression, be they developmental, physiological or environmental in nature. Conversely, the established association of reporter activity with a given cellular status of development and/or differentiation can be exploited to probe cell identity.

In an instance of the latter application three functional markers of *Arabidopsis* embryogenesis obtained by promoter trapping, *POLARIS* and *EXORDIUM* (Topping *et al.*, 1994), and *COLUMELLA*, were individually crossed to two mutants, including *gnom* (Mayer *et al.*, 1993; see Section 1.2.4.2), affected in embryo and seedling morphogenesis (Topping & Lindsey, 1997). *EXORDIUM* expresses *GUS* activity in the radicle and faintly in the cotyledons of embryos at later stages of development, and in cotyledons and shoot and root apices of seedlings (Topping *et al.*, 1994). *COLUMELLA* characteristically expresses *GUS* in the root cap of both primary and lateral roots (Topping & Lindsey, 1997). In homozygous *gnom* seeds and seedlings of *EXORDIUM* background *GUS* activity was found only in cotyledons, whereas in *COLUMELLA* background *GUS* expression was not observed. These findings were consistent with the view that *gnom* lacks both shoot and root meristems (Mayer *et al.*, 1993). In *POLARIS* embryos and seedlings *GUS* activity is restricted to the root tip (Topping *et al.*, 1994). Homozygous *gnom* mutants express the *POLARIS* marker in a polarized manner, even though they lack root meristems. This demonstrated *POLARIS* to be a marker of cellular position rather than structural identity. Furthermore, in severely affected (ball-shaped) *gnom* embryos in *POLARIS* background characterized by lack of cell structural differentiation, *GUS* activity is nevertheless polarized. Thus, the establishment of the apical-basal axis of polarity can be independent of structural differentiation.

1.3.5.5 Activation tagging

A gain-of-function mutagenic strategy has been devised which relies on the activation of native genes by an active promoter delivered to the genome by the interposon (Walden *et al.*, 1994). Mutations induced in this manner are expected to be phenotypically dominant, allowing their detection and the isolation of the affected gene. Fritze *et al.* (1995) used T-DNA-based activation tagging to isolate tobacco cell lines resistant to polyamine biosynthetic inhibitors. The tagged DNA of one of the selected lines conferred resistance to transformed protoplasts demonstrating its role in the acquisition of the trait.

The gain-of-function mutagenesis provided by activation tagging is relevant in functional genomics as the phenotype may provide important clues on the role of the activated gene.

1.4 Objectives of the research

1.4.1 Promoter-trapping and embryogenesis

The study of embryogenesis in plants can greatly benefit from the application of promoter-trapping both by its capacity to generate mutations causing phenotypic changes as well as to create functional markers, investigative tools of great utility (see above, Section 1.3.5.4). For the isolation and cloning of genes of interest promoter-trapping offers a number of advantages over other methods, not least of which is the ease of recovery of the tagged genes. Selection of the genes is not limited by a required association with a phenotypic trait as is for positional cloning, by the abundance of the products of expression or their ease of isolation, or by the availability of clones for homologous genes. In contrast, it benefits from the possibility to select genes based on the tissue-specificity of their expression, irrespective of its level, functional redundancy, or even lethality in the majority of cases when the altered traits are recessive. Several genes active during embryogenesis have been isolated and identified, but very few that are required during the early stages. Most of the exceptions are pattern mutants, like *gnom* (see Section 1.3.4.3), and only very recently their molecular characterization has been initiated. This is largely due to difficulties in accessing zygotic embryos and in obtaining sufficient amounts of material for manipulation. The creation of transcriptional and translational fusions with reporter genes like *gus* can greatly facilitate the identification of genes participating at all stages of the process. In addition, it permits the *in situ* study of changes in their expression.

In spite of the existence of numerous *Arabidopsis* transformants of promoter- or enhancer-traps (Clarke *et al.*, 1992; Devic *et al.*, 1995; Kertbundit *et al.*, 1991), their characterization and application to investigate embryogeny is still considerably limited. The largest publicly available collection of promoter-trap transformants, with over 2,000 lines, was produced by Clarke *et al.* (1992) with the T-DNA vector p Δ *gus*Bin19 where the promoterless

gusA gene is located in the left border region (Topping *et al.*, 1991). To exploit promoter-trap technology in the study of embryo development Lindsey and co-workers screened 430 of the transgenic lines in this library for GUS activity in siliques (Topping *et al.*, 1994). Of these, 74 (17.2%) exhibited GUS activity in siliques, with levels that varied over a 16-fold range. Thirty transformants with the highest levels of GUS expression were histochemically examined for reporter activity in developing embryos and six positives were identified. This indicated that GUS activity should be found in embryos of about 3-4% of transformants. In good agreement with this estimate Devic *et al.* (1995) found GUS activity in developing embryos of six transformants of a small collection of 100 produced with the same interposon. About 700 of the lines transformed with p Δ *gus*Bin19 have also been screened for mutants affected in aspects of embryogenesis including pattern formation or shape, and cytodifferentiation (Lindsey *et al.*, 1996). Several have been identified, occurring with a frequency of about 1-2%. Curiously, of the six embryo-GUS positives from the same library of transformants only two have been shown to have an inconspicuous associated phenotype (Topping *et al.*, 1994; Prof. K. Lindsey, personal communication). Of the shape mutants where it was examined, GUS activity was only detected in specific tissues of *gnom* but not in *rootless* and *hydra* (Lindsey *et al.*, 1996).

Although a number of transgenic lines were identified that are of interest for the study of embryogeny, their molecular characterization was very limited. Segregation analysis of the T-DNA selection marker (*nptII* gene) showed that only three of the six functional markers contained a single T-DNA locus (Topping *et al.*, 1994). Fusion transcripts were also detected in these three lines and left border junction fragments were cloned. For several of the visible mutants co-segregation of the phenotype with T-DNA was also demonstrated (Lindsey *et al.*, 1996). However, very little is known about the integration pattern of the p Δ *gus*Bin19 promoter-trap T-DNA. Activation of *gus* is known to occur in a significant proportion of transformants when different organs and developmental stages are considered. (50-60%; Prof. K. Lindsey, personal communication). For specific organs, reporter activity was exhibited in up to 30% of transformants in roots (Lindsey *et al.*, 1996). In their smaller set of lines Devic *et al.* (1995) found GUS activity in 62% of transformants. Using a distinct T-DNA promoter-trap (with *gus* located at the right border), Kertbundit *et al.* (1991) observed reporter activation resulting from transcriptional fusions in a comparable proportion of 54% of transgenic lines. Given the high ratio of activation generally observed and the many thousands of genes estimated to be expressed in the developing embryo (see Section 1.2.4.1), the low proportion of lines where reporter activity was registered in new sporophytes, is surprising. Also incongruent is the apparently low fraction of lines of transformants with an associated phenotype, but this might have result from limitations of the screening procedure and/or functional redundancy. It is then legitimate to enquire on the nature of the promoters driving the expression of the reporter.

The molecular characterization of functional markers can be greatly beneficial in several other respects. Tagged genes active during embryo formation can be isolated and, more

importantly, their functions identified. This would help to interpret the normal and altered patterns of expression of the associated markers. Also, the isolation of their promoters opens possibilities, so far not explored, to investigate the regulatory networks that control their activity. To date, molecular identification of *cis*-acting elements in promoters of embryo-active genes, and of interacting *trans*-acting factors has been limited to those highly expressed storage protein and *Lea* genes (reviewed in Thomas, 1993).

1.4.2 Specific objectives of the research

Within the above exposed context, the overall aims of the research with which this thesis is concerned were a detailed investigation of the T-DNA interposon p Δ *gus*Bin19 as a promoter trap and gene tagging element, and as an insertional mutagen, with particular reference to embryogenesis. Specifically, the research aimed at determining the correlation between the reporter gene activity in embryos of a transgenic line and the possible presence of a tagged native transcriptional unit. It also was directed at investigating the possible relation between the expression pattern of the reporter gene in the transgenic line and that of a putatively tagged gene in wildtype *Arabidopsis*. Line AtEN101, where GUS activity was limited to a specific developmental time window during embryogeny (see Section 3.2; Topping *et al.*, 1994) was selected for this study. In addition, the native promoter eventually detected in this line was to be delimited in preparation for its subsequent use in the eventual isolation of *trans*-acting factors that interact with it to activate or repress its activity. One of the recognized advantages of promoter-trapping over other methods of directed gene isolation is the possibility to identify functionally redundant genes (see Section 1.2.4.2). The absence of a visible phenotype in AtEN101 homozygous plants (Topping *et al.*, 1994) offered an opportunity to investigate the possible functional redundancy of an eventually tagged gene.

Additionally, to enrich the reduced number of transgenic lines with morphologically defective embryos (Lindsey *et al.*, 1996), a screening for additional mutants of this type was to be conducted. It was anticipated that the possible tagging of the affected genes with T-DNA from p Δ *gus*Bin19 would permit their quick isolation and characterization in future work.

Chapter 2

Materials and Methods

2.1 Materials

2.1.1 Chemicals and Reagents

2.1.1.1 General sources

With the exceptions indicated in the text, all bulk and fine chemicals and reagents used were of analytical reagent grade, and were purchased from either Amersham International, plc. (Little Chalfont), Fisher Scientific (Loughborough), Merck Ltd. (Poole), or Sigma Chemical Company Ltd. (Poole).

2.1.1.2 Water

Deionized distilled water (Q H₂O) was obtained from a Milli-Q Reagent Water Purification System (Millipore, Watford).

2.1.1.3 Preparation of phenol solutions

Liquified phenol pre-washed with Tris buffer (Fisher Scientific) was utilized unless otherwise indicated. When saturation with other buffers and/or change of pH was required the phenol was washed several times with the new buffer and the pH change monitored.

To determine the pH of a saturated phenol solution, 2 ml of the organic phase of the solution were mixed with 5 ml of methanol and 13 ml of Q H₂O. The pH of this solution was measured as normally with a pH meter probe.

2.1.2 Biochemicals

2.1.2.1 Radioisotopes

All radioisotopes were obtained from Amersham International plc (Little Chalfont). They included [α -³⁵S]dATP supplied at a concentration of 370 MBq/ml and with a specific activity greater than 22 TBq/mmol, and [α -³²P]dCTP at a concentration of 370 MBq/ml and with a specific activity of 110 TBq/mol.

2.1.2.2 Oligodeoxynucleotides

Oligodeoxynucleotides were either supplied by the Protein and Nucleic Acid Chemistry Laboratory (Leicester University), or by PE-Applied Biosystems UK (Warrington).

2.1.2.3 Enzymes, other proteins and molecular biology kits

Enzymes and other proteins

Restriction enzymes were from either Gibco/BRL Life Technologies Ltd. (Paisley) or from Promega UK (Southampton). The Klenow fragment of *E. coli* DNA polymerase I, T4 DNA ligase, and SuperScript™ II reverse transcriptase were supplied by Gibco/BRL. Exo⁻ Klenow was from Stratagene and DNAase-free RNAase from Amersham. RNAase-free DNAase, Proteinase K and calf intestine alkaline phosphatase were obtained from Boehringer Mannheim UK Ltd. (Lewes). RNasin® ribonuclease inhibitor was purchased from Promega. Taq DNA polymerase was from Promega, or, as BioTaq™, from Bioline (London). Expand™ High Fidelity PCR System was purchased from Boehringer-Mannheim.

Molecular biology kits

The sources for the various molecular biology kits were as follows:

- T7 Sequencing Kit, Pharmacia
- TA Cloning® Kit, Invitrogen BV (The Netherlands)
- Dynabeads mRNA Isolation Kit, Dynal UK Ltd. (Wirral)
- Bio-Rad Protein Kit, Bio-Rad (Hemel, Hempstead)
- PRISM™ Ready Reaction DyeDeoxy™ Terminator cycle Sequencing kit, Applied Biosystems Ltd. (Warrington)
- 5' AmpliFINDER Race Kit, Clontech Laboratories Inc. (Palo Alto, USA)
- Wizard DNA Clean-up System, Promega
- Prime-It II Random Primer Labeling Kit, Stratagene Ltd.
- Nuc-Trap® Probe Purification Columns (Stratagene)
- GeneClean II® Kit, Bio 101 (La Jolla, USA)
- High Pure PCR Product Purification Kit (Boehringer Mannheim).

2.1.2.4 Antibiotics and other biochemicals

Antibiotics including ampicillin, methicillin, streptomycin and kanamycin, and other biochemicals were obtained from Sigma Chemical Company Ltd. (Poole), unless otherwise stated. X-gal and IPTG were supplied by Calbiochem-Novabiochem (Nottingham) and X-gluc by Biosynth G (Staad, Switzerland).

Aqueous stock solutions of antibiotics were prepared, filter-sterilized and stored at -20°C until use.

2.1.2.5 Plasmids

General purpose cloning vectors:

For bacterial cloning of DNA fragments several vectors were used, including:

pUC18 (Yanisch-Perron *et al.*, 1985). This plasmid vector enables the histochemical detection of colonies of recombinants through the absence of α -complementation of the *lacZ* gene product in an appropriate host. The detection is based on absence of conversion of the chromogenic substrate X-gal to a blue-coloured product by β -galactosidase activity in bacteria harbouring a recombinant plasmid. Selection of pUC18 and derivatives relies on the expression of its ampicillin resistance gene (*amp*).

pBluescript (Short *et al.*, 1988). This is a family of phagemid vectors derived from the pUC plasmid series and containing the bacteriophage M13 origin of replication. They are distinguished solely by the relative orientation of the origins of replication and their polylinkers. Histochemical detection and positive selection with ampicillin are possible as with pUC18.

pCRTMII (Invitrogen). This vector is designed and was utilized for cloning of DNA products amplified with *Taq* DNA polymerase. It was obtained as part of the TA Cloning[®] Kit.

pGEMTM-T (Promega). As pCRTMII, this vector is designed and was utilized for cloning of DNA products amplified with *Taq* DNA polymerase.

Other plasmids

pGUS, a generous gift from Dr. J. F. Topping (University of Durham), is a derivative of pUC8 containing an *EcoR* I-*Bam*H I insert corresponding to a complete *gus* gene and the nopaline synthase transcription terminator.

pPRGUS-6 was constructed for this work and contains a single *EcoR* I-*Bam*H I insert derived from pGUS (see above) in pBluescript SK(+).

pMKC1 was a gift of Dr. M.-K. Cheung (University of Leicester). It includes a modified *gus* gene located downstream from the 35S CaMV promoter in a pBluescript backbone.

pTPRH29.E3 is a pUC18 derivative whose insert is a 2.4 kb *EcoR* I fragment of the rDNA unit of *B. rapa* (da Rocha & Bertrand, 1995). It comprises 0.1 kb of the 18S rRNA gene, the ITS spacer and 0.8 kb of the 25S rRNA gene.

EST clone 147C20T7 was obtained from the *Arabidopsis* Biological Research Center (Ohio State University, Columbus, Ohio).

2.1.2.6 Libraries of clones

A cDNA library prepared from *A. thaliana* poly(A)⁺ RNA isolated from siliques at various developmental stages was kindly supplied by Dr. J. Giraudat (Gif-sur-Yvette, France). The cloning vector had been λ Zap II (Stratagene). The inserts in the clones are flanked on each side by *EcoR* I sites derived from the polylinkers in the vector.

A library of genomic clones of *A. thaliana* was kindly supplied by Drs. Dangel (Köln) and Dr. Davis (Stanford University, U.S.A.). The library had been constructed in λ GEM11 vector (Promega) with DNA isolated from aseptically grown seedlings of the Columbia cultivar. Prior to cloning the genomic DNA fragments had been ligated to *EcoR* I adaptors.

2.1.3 Bacterial strains and media

2.1.3.1 Bacterial strains

Several strains of *E. coli* were used including a derivative of DH1 carrying R388::Tn1000XR and MH1578 (Sedgwick & Morgan, 1994), both generously offered by Dr. Sedgwick (London). The genotypes of the various strains are according to Sambrook *et al.* (1989), unless otherwise indicated:

- DH1 Tn1000XR derivative: F⁻, *recA1*, *endA1*, *gyrA96*, *hsdR17*, *supE44*, λ^- , R388::Tn1000XR (Sedgwick & Morgan, 1994).
- DH5 α (Bethesda Research Laboratories): F⁻, ϕ 80d*lacZ* Δ M15, Δ (*lacZYA-argF*)U169, *endA1*, *recA1*, *hsdR*, *gyrA96*, *supE44*, λ^- .
- INV α F': *endA1*, *recA1*, *hsdR17*(r^{-k+}, m⁺k), *supE44*, λ^- , Φ 80 *lacZ* α Δ M15 Δ (*lacZYA-argF*), F'
- LE392: F⁻, *hsdR514*, *supE44*, *supF58*, Δ (*lacIZY*)6, *trpR55*, *galK2*, *galT22*, *metB1*, λ^- .

- MH1578: streptomycin-resistant derivative of DH1 (Sedgwick & Morgan, 1994).
- SOLR (Stratagene).
- XL1-Blue: *recA1*, *endA1*, *gyrA96*, *hsdR17*, *supE44*, *relE44*, *lac*, {F', *proAB*, *lacI*^qΔM15, *Tn10*, (*tet*^R)}

2.1.3.2 Bacterial media

Solid media was prepared as described below for liquid media but bacto-agar was added to final concentration of 15 g/l. For top agar 7 g/l were used.

After their preparation, all bacterial media were sterilized by autoclaving at 121°C for 20 min.

Their compositions were as follows, per litre:

LB medium:	Bacto-tryptone	10 g
	Bacto-yeast extract	5 g
	NaCl	10 g
	(pH=7 by dropwise addition of 1 N NaOH)	

λ broth:	Bacto-tryptone	10 g
	NaCl	2.5 g
	Maltose	2%

(After autoclaving, the medium was made 10 mM in MgSO₄ by addition of filter-sterilized 2M stock).

SM:	NaCl	5.8 g	(final 0.1 M)
	MgSO ₄ .7H ₂ O	2.0 g	(final 8 mM)
	1 M Tris.HCl (pH 7.5)	50 ml	(final 50 mM)
	2% gelatin solution	5.0 ml	(final 0.01%)

2xYT:	Bacto-tryptone	16 g
	Bacto-yeast extract	10 g
	NaCl	5 g
(pH=7.0 by dropwise addition of 5 N NaOH)		

SOC:	Bacto-tryptone	20 g
	Bacto-yeast extract	5 g
	NaCl	0.5 g
	0.25 mM KCl	10 ml (final 2.5 mM)

(pH=7.0 with 5 N NaOH. Following sterilization by autoclaving the media was made 10 mM in MgCl_2 and 20 mM in glucose by addition of filter-sterilized 2 M MgCl_2 of 1 M of glucose solutions).

When necessary antibiotics were added from concentrated stocks (Section 2.1.2.4) to final concentrations of 50 $\mu\text{g/ml}$ for ampicillin and methycillin, and 100 $\mu\text{g/ml}$ for streptomycin. Molten solid medium was cooled to 50°C prior to their addition.

2.1.4 Plant material and growth

2.1.4.1 Plant Material

Two wildtype *A. thaliana* varieties, C24 and Columbia, were used. Seeds for the C24 variety were supplied by Dr. J. F. Topping, whereas seeds for the Columbia variety were a gift from Dr. Anna Sorensen (Leicester University).

Seeds for the lines transformed with p $\Delta\text{gusBin19}$ were supplied also by Dr. J. F. Topping. These included among others the T-DNA transgenic line AtEN101 and P26K4, later designated *bashful*. These lines were derived from C24 plants.

Seeds homozygous for the *diminuto* mutation were a generous gift from Prof. N.-H. Chua (Rockefeller University). This mutant had been isolated in C24 background but repeatedly backcrossed into the Columbia wildtype cultivar (Takahashi *et al.*, 1995).

Seeds for the *Nicotiana tabacum* cv. SR1 were a generous gift of Dr. John Gatehouse.

2.1.4.2 Plant culture media

1/2 MS: Half-strength Murashige and Skoog media (Sigma M5519), pH=5.7 (with dilute KOH).

1/2 MS₁₀: Half-strength Murashige and Skoog media (Sigma M5519) supplemented with 10 g/l sucrose, pH=5.7 (with dilute KOH).

MS₃₀: Full-strength Murashige and Skoog media (Sigma M5519) supplemented with 30 g/l sucrose, pH=5.7 (with dilute KOH).

To prepare solid media, agar was added to 8 g/l prior to sterilization. All media was autoclaved for 15 min at 121°C. When addition of kanamycin or of hormones was required, the molten media was pre-cooled to 50°C.

2.1.4.3 Hormone and antibiotic preparation and use

Kanamycin was prepared as an aqueous stock at 50 mg/ml, filter sterilized and used at concentrations of 25, 35 and 50 µg/ml in solid agar plates.

Absciscic acid (ABA; Sigma A-5199). Dissolved in 1 N NaOH.

2,4-Dichlorophenoxyacetic acid (2,4 D; Sigma D-8407). Dissolved in ethanol.

Epibrassinolide (BL; Sigma E-1641). Dissolved in DMSO.

Gibberillic Acid (GA₃; Sigma G-7645). Dissolved in ethanol.

Indole-3-Acetic Acid (IAA; Sigma I-2886). Dissolved in ethanol.

2.1.4.4 Seed sterilization, germination and growth

For germination and growth in soil the seeds were simply imbibed overnight in H₂O at 4°C, sown and left at that temperature in the dark for 5-7 days for vernalization and to ensure a uniform germination time.

For growth in media, *Arabidopsis* and tobacco seeds were sterilized in 10% sodium hypochlorite, 0.01% Tween 20 for 20 min. They were then washed several times with sterile H₂O sufficient to ensure a minimum 10⁵ dilution of residual sterilizing solution. The seeds were left imbibing in H₂O overnight at 4°C and then plated or placed in liquid medium.

For plating, seeds were usually individually picked with sterile forceps and placed on solid media in Petri dishes or sealed plastic pots. Petri dishes were sealed with Micropore™ (3M) tape. Plated seeds and liquid culture flasks were placed at 4°C in the dark for 5-7 days. They were then moved to growth rooms at 22-25°C, on day cycles of 16 h of white light with a photon flux density of 50-150 $\mu\text{mol}/\text{m}^2/\text{s}$, and 8 h of darkness. Liquid cultures were left shaking at low speed in orbital shakers.

For some experiments designed to assess the effects of hormones and other substances young seedlings were transferred from 1/2MS₁₀ or 1/2MS agar plates to liquid culture media supplemented as required. These seedlings were in some instances immersed in the liquid media contained in small Petri dishes and placed in orbital shakers. However, in most cases, to ensure that only their roots were immersed in the liquid medium, the seedlings were placed in polystyrene floats and transferred to sealed plastic pots.

Tobacco SR1 plants used for transient expression assays were grown in MS₃₀ in sealed plastic pots. Only relatively young plants 5-8 weeks old were used, as described in Section 2.9.

2.2 Bacterial methods

2.2.1 Growth and storage of bacterial strains

E. coli strains were normally grown in liquid LB medium at 37°C (Section 2.2.1), supplemented with antibiotics where required for maintenance of plasmids. Small liquid cultures (up to 50 ml) were initiated by inoculating with a single colony scraped from a plate. Larger cultures were initiated by inoculation from a small liquid culture grown overnight.

Bacteria were plated on solid agar plates of LB medium supplemented with antibiotics where required. Plating was performed by spreading a liquid culture with a glass rod or by streaking a colony with a sterile toothpick. For testing α -complementation, prior to spreading the bacteria 40 μl of X-gal (20 mg/ml in dimethylformamide) and 20 μl of IPTG (40 mg/ml in H₂O) were applied to the plate and allowed to dry for 30 min. The plated bacteria were incubated overnight at 37°C.

For storage, *E. coli* strains were grown to saturation in liquid LB medium (Section 2.1.3.2) supplemented with appropriate antibiotics when required. 0.7 ml of the bacterial culture were mixed with 0.5 ml of glycerol (sterilized by autoclaving) in a 2 ml cryo-tube and vortexed to mix. The suspension was quickly frozen in liquid nitrogen and stored at -70°C. New cultures

were initiated by scraping the top of the frozen stock with a sterile wire loop, streaking it on a plate of solid agar medium and incubating it overnight at 37°C.

2.2.2 Transformation of bacteria

Cell preparation

In most instances transformation of bacteria with plasmids or ligation mixes was performed by electroporation, according to Dower *et al.* (1988). Either XL1-Blue or, more commonly, DH5 α were used as hosts. To prepare electroporation-competent cells a small (20 ml) culture was grown overnight in LB at 37°C. 1 ml of this culture was inoculated into each of two 200 ml of pre-warmed LB medium in 1 liter flasks. The new cultures were grown with vigorous shaking until OD₆₀₀=0.5-0.7. The cultures were quickly chilled on ice and the cells pelleted by centrifuging at 4,000 x g for 15 min at 4°C. The pelleted cells were kept cold (on ice) and very gently resuspended in 200 ml of ice-cold 1mM HEPES, pH 7.0 (sterilized by filtration). The suspension was centrifuged and the cells resuspended as before. After pelleting the cells by identical centrifugation, they were resuspended in 20 ml of ice-cold 10% glycerol (sterilized by autoclaving). Cells were pelleted once more and resuspended in 1-2 ml of ice-cold 10% glycerol. This concentrated suspension was dispensed in 50-100 μ l aliquots into pre-cooled minifuge tubes and frozen by immersion into liquid nitrogen and stored at -70°C until use.

Cell transformation

A pulse generator was set to 25 μ F capacitor, 2.5 kV and 200 Ω in parallel with the sample chamber. Clean 200 μ l electroporation cuvettes and the cuvette holder were kept cold on ice.

Aliquots of the frozen cells were thawed at room temperature and placed on ice. 40 μ l of the cell suspension were gently pipetted into a pre-cooled minifuge tube. 1-2 μ l of DNA solution in a low-ionic strength solvent (Q H₂O or TE buffer) were gently mixed in and the suspension left on ice for about 1 min.

The cell suspension was transferred to the bottom of an electroporation cuvette. The cuvette was placed in the pulse generator and the pulse applied. 1 ml of SOC medium was immediately added to the cuvette and mixed with the cell suspension. The mix was transferred to a 20 ml sterile plastic tube and incubated with shaking at 37°C for 1 h. Afterwards, the suspension was plated in selective LB solid agar medium. The plates were incubated overnight at 37°C.

2.2.3 Generation of transposon-mutagenized plasmids

To generate a collection of derivatives of plasmid pENB8 (Section 8.2.1) insertional modification by random integration of the transposon Tn1000XR (Sedgwick & Morgan, 1994) was used. The procedure followed the recommendations of Professor Sedgwick (1994; personal communication). The transposon donor bacterial strain was a derivative of DH1 harbouring R388::Tn1000XR, and the recipient strain was MH1578, a streptomycin-resistant derivative of DH1 (Section 2.1.3.1).

The plasmid to be mutagenized was transformed into the donor bacteria as described (see above, Section 2.2.2). Selection of the transformants utilized ampicillin and methicillin. 10 ml cultures of the transformed donor strain in LB plus ampicillin-methicillin, and of the recipient strain MH1578 in LB were grown overnight. From each, 0.1 ml were used to initiate fresh cultures in 10 ml of the same media. The cultures were allowed to grow until log phase ($OD_{600}=0.2$). About 3 ml of the donor culture were pelleted in a minifuge, gently washed with fresh LB, re-pelleted and then resuspended in 1 ml of the MH1578 culture. This mating mixture was plated onto an LB agar plate and incubated at 37°C for 1 h. The mating plate was then rinsed with 10 ml of filter-sterilized 10 mM $MgSO_4$ and the cells collected from this suspension by centrifugation. The pellet was resuspended in 1 ml of LB and aliquots of 10-100 μ l were spread on LB plates supplemented with methicillin, ampicillin and streptomycin. The plates were incubated overnight at 37°C and the following day individual colonies were picked. Selection of clones was performed by preparing plasmids from individual colonies (Section 2.4.1.1) and analyzing their structure.

2.3 Bacteriophage methods

2.3.1 Plating of λ phage

Plating of λ clones was used to titer the phage suspensions, screen libraries and to generate new stocks.

Solutions

- 0.2% maltose (filter-sterilized)
- 10mM $MgSO_4$ (filter-sterilized)
- SM
- λ top agar
- Top agar (molten, kept at 47°C in a water bath)
- LB plates (warmed to 37°C)

2.3.1.1 Titering

A colony of LE392 was inoculated into 50 ml of λ broth. The culture was grown overnight at 37°C. The OD₆₀₀ of the culture was determined (after diluting 1:4 in the same media).

LE392 cells were pelleted by centrifuging in sterile plastic tubes at 4,000 x g for 10 min at RT. The pellet was resuspended in 10 mM MgSO₄ solution to an OD₆₀₀=3.0. The cell suspension was immediately used or stored at 4°C until use.

Ten-fold serial dilutions of bacteriophage(s) to be titered (or to re-stock) were prepared in SM. 0.1 ml aliquots of the cell suspension were dispensed into 20 ml sterile plastic tubes. To each 0.1 ml of a phage dilution was added and mixed in by quick vortexing. The tubes were incubated at 39°C for 20 min.

3 ml of molten top agar were added to each tube quickly mixed in by low-speed vortexing. The contents were immediately poured onto a pre-warmed LB plate. The covered plates were incubated at 37°C until the plaques could be counted (usually about 16 hours).

2.3.1.2 Preparation of λ stocks

Selected plaques of λ phage were picked by piercing through the agar with a sterile pasteur pipette, sucking in the agar containing the plaque and dispensing it into a sterile storage tube containing 1 ml of SM and a drop of chloroform. When picking from plated libraries sometimes the desired plaque could not be precisely located and it was necessary to carve out several with a sterile blade. The phages were allowed to diffuse into the SM for 1 hour at RT and then stored at 4°C.

2.3.2 Screening of libraries in λ phage

Reagents and solutions

Denaturing solution (0.5 M NaOH, 1.5 M NaCl)

Neutralizing solution (1.5 M NaCl, 0.5 M Tris.HCl pH 7.4)

2 x SSC (0.3 M NaCl, 0.03 M trisodium citrate)

Primary screening

Plating bacteria were prepared as described above (2.3.1.1) and 2.0 ml aliquots were dispensed into 50 ml sterile plastic tubes. An estimated 80,000 to 120,000 pfu of the library to be screened in less than 50 μ l were mixed with each aliquot of bacteria. The suspensions were incubated at 39°C for 20 min.

After infection, 30.0 ml of molten top agar were mixed with the bacterial suspension and quickly mixed by vortexing. The mix was immediately poured onto a large 240 x 240 mm bottom agar plate.

The plates were incubated at 37°C until the plaques had the appropriate size. The plates were cooled to 4°C to slow plaque growth and harden the agar.

On top of the agar in each plate a nylon membrane filter (Bio-Rad) previously cut to a suitable size was layed for 1 min. To orient the filter relative to the plate, a needle filled with ink was stabbed through the filter and the agar at various points. The filter was removed and treated as described below. Duplicate and triplicate filters were prepared by laying them over the plates for 2-3 min (5 min for triplicates). The filters were oriented by stabbing at the previously pierced positions visible in the agar.

After their removal from the agar plates each filter was placed with the DNA side up on 3 MM Whatman filter paper soaked in denaturing solution for 5 min. It was then transferred to a 3 MM Whatman paper soaked in neutralizing solution. After 5 min it was removed and layed onto 3MM Whatman paper saturated with 2 x SSC. At the end of 5 min the nylon filter was left to dry on filter paper, with the DNA side up.

When all filters were dry, they were each sandwiched between two sheets of 3 MM Whatman filter paper and baked for 2 hours at 80°C in a vacuum oven. The nylon membranes were then stored until further use.

The nylon membranes were hybridized with appropriate probes as described in Section 2.8.3 and autoradiographed (Section 2.8.5). Hybridizing plaques were isolated from the plates and λ stocks prepared from them (Section 2.3.1.2).

Secondary and tertiary screens

For all clones secondary and tertiary screens were performed to confirm the sequence similarity to the probes and eliminate other contaminating phage. These screens were conducted as described above except that proportionally smaller amounts of serially diluted phage in SM were mixed with 0.3 ml of plating bacteria and plated on 9 cm Petri dishes after dispersion in 3 ml of top agar.

2.3.3 Excision of phagemid from λ ZAPII

To excise the pBluescript phagemid in individual λ ZAPII clones the Ex-Assist/SOLR system (Stratagene) was used. The protocol recommended by the manufacturer was closely followed.

Overnight 5 ml cultures of *E. coli* strains XL1-Blue and SOLR were grown in λ broth and LB, respectively. The XL1-blue culture was diluted to $OD_{600}=1.0$ with λ broth and 200 μ l dispensed into a 25 ml plastic tube.

To the XL1-Blue cells in the plastic tube 100 μ l of the λ clone from an SM stock were added together with 1 ml of Ex-Assist helper phage ($> 1 \times 10^6$ pfu/ μ l). The mix was incubated at 37°C for 15 min. 3 ml of 2xYT were then added and the suspension incubated at 37°C for 2 h.

An aliquot of the suspension was transferred to a microfuge tube and warmed in a water bath at 70°C for 20 min. The dead cells were removed by centrifugation at 12,000 x g for 2 min at RT. The supernatant containing the phagemid in the form of filamentous phage was transferred to a clean minifuge tube and stored at 4°C.

E. coli clones of the plasmid were obtained by mixing 50 μ l of the phage with 200 μ l of SOLR cells (previously diluted to $OD_{600}=1.0$) and incubating at 37°C for 15 min. Ten-fold serial dilutions of the infected cells were prepared and then plated out on LB agar plates supplemented with ampicillin (50 μ g/ml). Individual colonies were picked from the plates after overnight incubation at 37°C.

2.4 DNA isolation

2.4.1 Plasmid DNA isolation

2.4.1.1 Mini-preparations

Reagents

Solution I (50 mM glucose, 25 mM Tris.HCl, pH 8.0, 10 mM EDTA, pH 8.0).

(This solution was autoclaved and stored at 4°C)

Solution II (0.2 N NaOH, 1% SDS)

Solution III (stored at 4°C). Prepared by mixing of:

5M Potassium acetate	60.0 ml
Glacial acetic acid	11.5 ml
Q H ₂ O	28.5 ml

From a freshly grown colony of an *E. coli* transformant of interest a small inoculum was scraped with a sterile toothpick. The inoculum was transferred to 3-10 ml of liquid LB media supplemented with the required antibiotics (usually ampicillin and methicillin) in a 15-50 ml tube. The new culture was allowed to grow overnight at 37°C.

One (or more) 1.5 ml aliquot of cell suspension was transferred to a microfuge tube and spun at 12,000 x g for 1 min. The supernatant was removed and the cells were washed by resuspending in 1.0 ml of ice-cold Solution I and pelleting as before.

After removing the supernatant, the cells were resuspended in 100 µl of ice-cold Solution I. Then 200 µl of freshly prepared Solution II were added to the tube to lyse the cells. The contents were mixed by inversion and left to rest on ice for 1-3 min.

Proteins and nucleoid DNA were precipitated by addition of 150 µl of ice-cold Solution III followed by vigorous mixing. The solution was kept on ice for 5 min and then spun at 12,000 x g for 5 min at 4°C.

The supernatant containing the plasmid DNA was transferred to a new minifuge tube and extracted with 0.4 ml of phenol/Sevag (1:1). The phases were separated by centrifuging at 12,000 x g for 5 min at 4°C.

The supernatant was recovered and re-extracted with 0.4 ml Sevag to remove traces of phenol. The phases were separated as before.

The recovered supernatant was mixed with two volumes of ethanol to precipitate the plasmid. After leaving for at least 5 min, the DNA was pelleted by centrifuging at 12,000 x g for 5 min at 4°C. The pellet was washed twice with 70% ice-cold ethanol and resuspended in 20-50 µl of TE buffer containing DNAase-free pancreatic RNAase (20 µg/ml). To remove RNA, the solution was incubated at 37°C for 30 min, and then stored at -20°C.

2.4.1.2 Large-scale preparations

Reagents

Solutions I, II and III as above, Section 2.4.1.1

8M LiCl

PEG/NaCl solution (1.6 M NaCl; 13% PEG 6000)

5-10 ml of an overnight culture of the bacterial clone of interest grown in LB supplemented with the required antibiotics were inoculated into a larger volume (400 ml) of the same medium. This culture was incubated at 37°C overnight with shaking.

The culture was cooled on ice and then the cells were pelleted by centrifuging in 200 ml bottles at 5,000 x g at RT for 10 min.

To wash the cells, the pellets were resuspended in 20 ml of ice-cold Solution I and transferred to a plastic tube (50 ml). The cells were re-pelleted by centrifuging at 5,000 x g for 10 min at 4°C.

After decanting the supernatant, the pellet was thoroughly resuspended in 8 ml of ice-cold Solution I. For lysis, 12 ml of Solution II were added and mixed in by inverting the tube 4-6 times. Almost immediately 10 ml of Solution III were added to precipitate proteins and chromosomal DNA. Mixing was by inverting the tube 6-10 times. The tube was kept on ice for at least 15 min.

The tube was centrifuged at 12,000 x g for 5 min at 4°C. The supernatant containing the plasmid was filtered through miracloth into a clean tube. To precipitate the nucleic acids 1 vol. of isopropanol was mixed in and the solution left to stand until precipitate became visible (often just a few min).

The precipitate was recovered by centrifuging at 12,000 x g for 10 min at 4°C. The pellet was completely resuspended in 3 ml of Q H₂O. The same volume of LiCl solution was added and well mixed in. The precipitate formed was removed by centrifuging at 12,000 x g for 10 min at 4°C.

To the supernatant, 1 vol. of isopropanol was added and mixed. The plasmid DNA was recovered by centrifuging at 12,000 x g for 10 min at 4°C. The plasmid DNA was dissolved in

500 µl of TE and mixed with 20 µl of DNAase-free RNAase A (20 mg/ml). After transferring to a microfuge tube, the solution was incubated at 37°C for 1 h.

To recover the plasmid DNA 500 µl of PEG/NaCl solution were mixed in and the solution was left on ice for at least 1 hour. The precipitated DNA was pelleted by centrifugation at 12,000 x g for 10 min at 4°C.

The DNA was re-dissolved in 300 µl of TE and extracted with an equal volume of phenol/Sevag (1:1). After resolving the phases by centrifuging at 12,000 x g for 5 min, the aqueous solution was transferred to a new tube. The phenol/Sevag extraction was repeated followed by a single extration with an equal volume of Sevag.

After recovering the aqueous phase the plasmid DNA was precipitated with the mixing in of 100 µl of 5M ammonium acetate and the addition of 2 volumes of ethanol. The DNA was recovered by centrifuging at the lowest speed possible or collected with a hook. Excess salt was removed by washing with 70% ethanol. The DNA was then re-dissolved in TE buffer

2.4.2 λ phage DNA

2.4.2.1 Large-scale isolation

Essentially performed as indicated by Sambrook *et al.* (1989).

A 10 ml culture of *E. coli* strain LE392 in λ broth was grown overnight at 37°C. 1-2 ml of this culture were used to inoculate two 500 ml volumes of the same medium each in 2 l flasks pre-warmed to 37°C. The cultures were grown with vigorous shaking until the OD₆₀₀ was 0.7 (it had been established that at this OD₆₀₀ the concentration of LE392 cells is about 1.7-2.0 x 10⁸/ml).

To each flask 10¹⁰ pfus of bacteriophage were added and incubation allowed to proceed until lysis occurred (usually from 5-8 hours). Lysis was accompanied by a sharp drop in OD₆₀₀ (*e.g.* from 2.0 to 0.4). At this point 10 ml of chloroform were added and the incubation allowed to proceed for 10 min to complete the lysis.

The lysed cultures were chilled to RT and pancreatic DNAase I and RNAase A were added to a final concentration of 1 µg/ml. The cultures were allowed to rest for 30-60 min at RT.

Solid NaCl (29.2 g) was dissolved in each suspension to a final concentration of 1 M. The suspension was cooled on ice for 1 h. Then, the debris was removed by centrifuging in 250 ml bottles at 11,000 x g for 10 min, 4°C.

The clean supernatants were decanted into 1l flasks and 10% (w/v) of solid polyethylene glycol 8000 were added and dissolved by slow stirring at RT. The suspension was allowed to rest for 1 h to overnight on ice.

To recover the bacteriophage particles the suspension was transferred to 250 ml centrifuge bottles and spun at 11,000 x g for 10 min at 4°C. The supernatant was discarded and the remainder was removed as thoroughly as possible with a pipette.

The phage particles were gently resuspended in 8 ml of SM with the help of a wide-bore pasteur pipette. PEG and cell debris was removed by extracting with 1 vol. of chloroform and vortexing briefly. The phases were separated by centrifuging at 3,000 x g for 15 min at 4°C.

The aqueous phase was then ultracentrifuged in an L-70 Beckman Ultracentrifuge at 25,000 rpm for 2 h, 4°C using a SW28 rotor. The supernatant was removed and 1-2 ml of SM were added to the tube and left overnight at 4°C. Resuspension was completed by pipetting with a wide-bore pasteur pipette. Proteinase K and SDS were added to final concentrations of 50 µg/ml and 0.5% respectively. The solution was incubated at 56°C for 1 h.

The solution was cooled to RT and extracted with 1 vol. of phenol. Following centrifugation at 3,000 x g for 5 min the aqueous phase was transferred to a clean tube and re-extracted with phenol/chloroform (1:1). A final extraction was performed with chloroform. The aqueous phase was dialysed against 1-2 liter of TE for at least 3 hours, and then against a new volume of TE overnight. The DNA solution was transferred to a clean tube and kept at -20°C until use.

2.4.2.2 Small-scale λ DNA isolation

According to Ausubel *et al.* (1996).

A 10 ml culture of *E. coli* strain LE392 in λ broth was grown overnight at 37°C. In a minifuge tube 100 µl of the culture were mixed with 100 µl of phage eluate (see Section 2.3.1.2) and 100 µl of the MgCl₂/CaCl₂ solution. The suspension was incubated in a water bath at 39°C for 15 min.

50 ml of pre-warmed λ broth medium were inoculated with the mix and incubated at 37°C with vigorous shaking until lysis (usually 6-8 hours). Upon lysis, about 0.5 ml chloroform were added and the culture shaken for another 5 min.

The suspension was transferred to 50 ml centrifuge tubes and spun at 12,000 x g for 10 min at 4°C. The lysate was then transferred to a new tube and pancreatic DNAase I and RNAase were added to 0.2 µg/ml and 5 µg/ml, respectively. The suspension was incubated for 1 h at 37°C.

The suspension was transferred to ultracentrifuge tubes and spun for 90 min at 27,000 rpm in a SW-28 rotor (132,000 x g) at 4°C. The supernatants were discarded and removed as thoroughly as possible.

The pellets were resuspended in 200 µl of 0.05 M Tris.HCl, pH 8.0. The suspension was transferred to a minifuge tube and then extracted with 200 µl of phenol by vortexing for 20 min. The phases were separated by centrifuging 5 min at 12,000 x g. The top aqueous phase was transferred to a new tube and re-extracted with phenol.

After phase separation as before, the aqueous suspension was extracted twice with an equal volume of chloroform by shaking. Following centrifugation, the top aqueous phase was transferred to a new minifuge tube and 1/10 vol. of 3M sodium acetate, pH 4.8 were mixed in. The DNA was then precipitated by addition of 2 vol. of ethanol.

The precipitated DNA was pelleted by centrifugation at 12,000 x g for 10 min. The supernatant was removed and the pellet washed twice with 1 ml of 70% ethanol. The DNA was dissolved in 50-100 µl of TE buffer.

2.4.3 Plant DNA isolation

2.4.3.1 CTAB method

This method is a modification of that of Murray and Thompson (1980) by Ausubel *et al.* (1996).

Reagents and solutions

CTAB extraction solution (2% (w/v) CTAB, 100 mM Tris.HCl, pH 8.0, 20 mM EDTA, pH 8.0, 1.4 M NaCl, 2% (v/v), β -mercaptoethanol)

The β -mercaptoethanol was added immediately before use.

CTAB/NaCl solution (10% (w/v) CTAB, 0.7 M NaCl)

Preparation of this solution requires heating.

CTAB precipitation solution (1% (w/v) CTAB, 50 mM Tris.HCl, pH 8.0, 10 mM EDTA, pH 8.0)

High-salt TE buffer (10 mM Tris.HCl, pH 8.0, 0.1 mM EDTA, pH 8.0, 1 M NaCl)

Procedure

Plant material (5-10 g fresh weight of leaves or seedlings) were pulverized while frozen with liquid nitrogen by grinding in a mortar, sometimes with the aid of a small amount of acid-treated white quartz sand.

The frozen powder was quickly added and gently mixed with CTAB extraction solution (1 ml:1 g) pre-heated to 67°C in a 50 ml plastic tube. The homogenate was incubated at 67°C for 1 h with occasional inversion.

The solution was cooled to room temperature and gently well mixed with a 1.0 volume of Sevag and left at RT for 15 min with occasional inversion. The phases were clearly separated by spinning at 4,000 x g for 5 min at RT.

The top aqueous phase was recovered and gently mixed with 1/10 vol. of pre-heated (67°C) CTAB/NaCl solution. The solution was re-extracted with 1 vol. of Sevag and the phases separated as before.

The aqueous phase was recovered and gently mixed with exactly 1 vol. of CTAB precipitation solution. The solution was left undisturbed for several hours (usually overnight) at RT. When no or little precipitate was visible up to 1/10 vol. of CTAB precipitation solution was gently mixed. The new solution was left to rest for several hours at 37°C.

The precipitate containing the DNA was pelleted by spinning at 1,000 x g at 4°C for 5 min. The supernatant was carefully and thoroughly removed. The pellet was re-dissolved in 0.5-1.0 ml of high-salt TE buffer, sometimes helped by heating at 65°C.

The nucleic acids were precipitated by the addition of 0.6 vol. of isopropanol. After mixing well the solution was spun at 10,000 x g for 15 min. at 4°C. The supernatant was discarded and the pellet rinsed twice with 80% ethanol.

The pellet of nucleic acids was resuspended in a minimal volume of TE. About 3-10 µl of DNAase-free RNAase (Amersham, 20 units/µl) were added and the solution was incubated at 37°C for 30 min.

2.4.4 Purification of DNA with Wizard™ clean-up system

For several purposes, especially for sequencing, previously isolated plasmid DNA (Section 2.4.1) was purified to a higher degree with the help of Wizard™ clean-up system minicolumns (Promega), following the instructions of the manufacturer. The DNA (up to 20-40 µg) in 100 µl of TE buffer was mixed with 1 ml of Wizard DNA Clean-Up Resin in a minifuge tube and mixed by inversion. The slurry was applied to a syringe barrel attached to the minicolumn and then slowly pushed through with the plunger. The trapped resin was washed with 2 ml of 80% isopropanol and centrifuged at 12,000 x g for 20 seconds to dry the resin. The remaining isopropanol was air dried for 5 min. 50 µl of pre-warmed TE buffer were then applied to the column. After 1 min the buffer was eluted into a minifuge tube to which the column was attached by centrifugation at 12,000 x g for 20 secs. The elution was repeated with another aliquot of 50 µl of TE. The concentration of DNA was estimated by UV absorption. If necessary, the DNA in the eluate was concentrated by ethanol precipitation. The DNA was stored at -20°C until use.

2.4.5 Purification of oligodeoxynucleotides

For removal of ammonia and other contaminants present in supplied oligodeoxynucleotides the procedure of Sawadogo and van Dyke (1991) was followed. 100 µl of oligonucleotide solution was vortexed vigorously in a microfuge tube with 1 ml of *n*-butanol for 15 sec. The tube was centrifuged for 1 min at 14,000 x g. The oligonucleotide pellet was extracted once again with *n*-butanol, dried under vacuum and resuspended in TE or Q H₂O.

2.4.6 Purification of DNA from gels

To purify DNA from bands in agarose gels (Section 2.5.2) two different approaches were used including electroelution and purification with a silica matrix (GeneClean II® Kit, Bio 101)

2.4.6.1 Electroelution

Preparation of dialysis tubing

Dialysis tubing (Sigma D-0405, cutoff > 12 kD) was cut into segments of about 20-30 cm and boiled for 10 min in 1-2 l of 2% sodium bicarbonate, 1 mM EDTA (pH 8.0). The tube was rinsed in Q H₂O and boiled once again for 10 min in 1 mM of EDTA (pH 8.0). The membrane was washed several times in Q H₂O and then stored in 50% ethanol at 4°C.

Procedure

The bands of interest were cut out of the gel with a razor blade under low-intensity U.V. light and placed onto a piece of Whatman[®] Laboratory Sealing Film. The gel fragments were rinsed with Q H₂O and then placed inside a pre-conditioned dialysis bag of the smallest size possible. The bag was filled with 1x TAE buffer (Section 2.5.2) and closed. Electroelution was conducted at up to 5V/cm and after 30-60 min the polarity of electrophoresis was reversed for 5-10 secs. The TAE inside the dialysis bag was recovered into minifuge tube(s) and extracted once with 1 vol. phenol/Sevag (1:1). After separation of the phases by centrifugation at 14,000 x g for 5 min, the aqueous phase was re-extracted with 1 vol. of Sevag. The solution was centrifuged as before and the aqueous phase was mixed with 1/10 vol. of 3 M sodium acetate (pH 7) and 2.5 vol. of ethanol. The DNA was left to precipitate at -20°C for at least 1 h to overnight. The DNA was pelleted by centrifugation for 20 min at 14,000 x g, 4°C, rinsed with 70% ethanol, and re-dissolved in TE buffer.

2.4.6.2 Purification with GeneClean II[®]

A block of agarose containing the band of fragments to be purified was excised from the gel and placed into a pre-weighted minifuge tube. Its weight was determined and two volumes of NaI stock solution (provided with the kit) were added and the tube warmed to 55°C until the agarose melted (5-10 min).

To the molten agarose a suspension of silica matrix GLASSMILK[®] was mixed in (about 1 µl/µg), and the tube cooled on ice for 5-10 min with occasional mixing. The silica matrix was then pelleted by centrifugation in a minifuge at top speed for 5 sec. The supernatant was removed as thoroughly as possible. The pellet was washed three times in about 500 µl of supplied New Wash solution. After each resuspension in New Wash the silica matrix was pelleted by centrifugation for 5 sec.

To elute the DNA the silica particles were resuspended in a variable volume of TE buffer and the tube incubated to 55°C for 2-3 min. The silica was then pelleted out by centrifuging at 14,000 x g for 1-2 min and the supernatant transferred to a new tube. Sometimes the elution step was repeated and the supernatants combined. The purified DNA was directly used or concentrated by precipitation with sodium acetate and ethanol.

2.4.7 Purification of PCR products

When required PCR products were purified by selective adsorption to a silica matrix. A High Pure PCR Product Purification Kit (Boehringer Mannheim) was used to this purpose following the instructions provided with some modifications.

To 100 µl of PCR reaction 500 µl of Binding Buffer (provided with the kit) were added and mixed. The sample was pipetted into a High Pure Filter placed on a Collection Tube. This assemblage was spun at 14,000 x g for 30 seconds in a minifuge. The flowthrough was

discarded and 500 µl of Wash Buffer added to the High Pure Filter. Washing was performed by centrifugation as before. A second wash with 200 µl of Wash buffer was conducted in a similar manner. The Collection Tube was discarded and the Filter unit placed on a clean minifuge tube. The DNA was eluted from the Filter with 50 µl of Elution Buffer by centrifugation at 14,000 x g for 30 seconds. The elution procedure was repeated with a second 50 µl aliquot of Elution Buffer. If necessary the DNA was concentrated by precipitation with sodium acetate and ethanol.

2.5 Manipulation and analysis of DNA

2.5.1 Enzymatic manipulation

2.5.1.1 Restriction endonuclease digestions

Digestions of plasmid DNA, genomic DNA or PCR products were performed with purified material in the buffers supplied by the manufacturer and under the conditions recommended for each enzyme. Normally 5 to 10 U of enzyme were used per µg of DNA. Digestions were for 3-5 hours except with genomic DNA when they were carried out overnight in the presence of spermidine trihydrochloride (0.4 mM). Digestions with multiple enzymes were performed in a commonly compatible buffer or sequentially by adjusting the buffer conditions or following a DNA purification step.

2.5.1.2 Filling-in of recessive ends

When necessary, filling-in reactions were performed to compatibilize 3' recessed ends prior to ligation. To the digested DNA sample in 20 µl reaction volume, 1 µl of each of the required dNTPs (20 mM stocks) were added. The reaction was initiated by the addition of 1 µl of 1 U/µl Klenow fragment of *E. coli* DNA polymerase I (Gibco/BRL). The reaction was allowed to proceed at RT for 15 min. The enzymes were removed and the DNA purified by the common procedure of phenol/Sevag extraction and sodium acetate/ethanol precipitation.

2.5.1.3 Dephosphorylation of 5' ends

In some instances, prior to ligation, the 5' protruding or blunt ends of one of the fragments were dephosphorylated to prevent self-ligation. For dephosphorylation, at the end of restriction enzyme digestion (in 25 µl) the enzyme was heat-killed at 70-75 °C, and 1 µl of ZnCl₂ (25 mM) was added to a final concentration 1mM. The reaction was initiated by addition of 1 U of calf intestinal phosphatase (CIP; 1U/µl) and continued at 37°C for 30 min. For blunt ends another 1 U of CIP was added and the solution incubated for a further 45 min at 55°C.

At the end of the CIP treatment, the reaction was stopped by the addition of 2 μ l of 100 mM EDTA (pH 8.0) and heating to 70°C for 30 min. Then the DNA was purified and CIP removed by the common procedure of phenol/Sevag extraction and sodium acetate/ethanol precipitation.

2.5.1.4 Ligation of DNA fragments

For the construction of recombinant plasmids linearized vector and insert fragments with compatible ends were ligated together with T4 DNA ligase (Gibco/BRL). Usually the reactions were conducted in a 15 μ l volume prepared from 3 μ l of supplied 5x ligase buffer and containing about 20-100 ng of vector DNA mixed the fragment(s) to be cloned in 1:1 up to 1:3 molar ratio, and 1 unit of T4 ligase. The reactions were allowed to proceed at room temperature for up to 5 min when the DNA termini were staggered. For blunt-end ligations the incubation temperature was 14°C and the reaction was allowed to proceed overnight. At the end of the incubation period the ligated DNA was purified by phenol:Sevag extraction followed by another extraction with Sevag. To the aqueous phase containing the DNA 1 μ l of glycogen solution (1 mg/ml) was added, 1/10 vol. of 3M sodium acetate was mixed in, and the DNA precipitated by the addition of 2.5 vol. of ethanol.

Prior to use in transformation the DNA was pelleted by centrifugation, washed with 70% ethanol, dried and resuspended in TE buffer.

2.5.2 DNA agarose gel electrophoresis

Solutions:

1x TAE buffer:

40 mM	Tris.Acetate, pH 8.0
1 mM	EDTA, pH 8.0
0.05 μ g/ml	Ethidium bromide

10x DNA loading buffer:

0.25% (w/v)	Bromophenol blue
0.25% (w/v)	Xylene cyanol FF
25% (w/v)	Ficoll (type 400)

Agarose gels of appropriate concentration for the size of the fragments to be separated (0.7-1.5%) were prepared in 1x TAE buffer by heating to boil in a microwave oven. Once the gel was homogeneously liquified the mix was cooled to about 50°C, supplemented with ethidium bromide to 0.10 μ g/ml, and poured onto a gel setting plate. After solidified by leaving to rest at

room temperature the gel was placed in the electrophoresis tank. The tank was filled with 1x TAE buffer supplemented with ethidium bromide to 0.1 µg/ml.

To help uniformizing the ionic strength of the various samples, occasionally 1/10 vol. of a concentrated (10x) restriction enzyme buffer was added to some (*e.g.* DNA markers). To all samples 1/10 vol. of 10x DNA loading buffer was mixed in. The solutions were heated at 65°C for 3-5 min in a dry bath and immediately loaded on the gel. Gels were run at 1-5 V/cm. To estimate the sizes of DNA fragments a DNA size marker was run in parallel. The marker used in most instances was the 1 kb DNA ladder (Gibco/BRL). A known amount of the marker was loaded which, when required, enabled an estimation of the DNA amounts in selected bands.

After electrophoretic resolution, the gels were rinsed in Q H₂O and placed over a UV transilluminator to visualize the DNA and photograph or video-record the results.

2.5.3 Southern blot preparation

For detection of electrophoretically resolved DNA fragments by hybridization with a labeled probe (Section 2.8.3), the DNA in gels was transferred by capillarity to a nylon membrane (Southern blotting).

After completion of electrophoresis the gel was immersed for 10-15 min in depurinating solution (0.25 M HCl). This step was omitted if the DNA fragments were less than 3 kb. The gel was subsequently placed on top of the transfer solution source. This source consisted of several layers of a smooth sponge topped with 3 layers of 3 MM Whatman filter paper all soaked with transfer solution (0.4 M NaOH), placed in a container of the same solution. A nylon membrane, Zeta-Probe[®] GT (Bio-Rad), of the same size as the gel (pre-treated by soaking in Q H₂O for 5 min) was placed over it. Three layers of 3 MM Whatman filter paper soaked in transfer solution were laid on top followed by a stack of dry paper towels. Capillary transfer was allowed to proceed for at least 5 h to overnight.

After blotting, the position of the loading wells were marked on the nylon membrane. The membrane was removed, briefly rinsed in 2x SSC and allowed to air dry on a filter paper. To fix the DNA the nylon membrane was baked in a vacuum oven for 2 hours at 80°C. The baked nylon membrane was stored in a dry place until use.

2.5.4 Sequencing of DNA

For sequencing plasmids, DNA was isolated as described in Section 2.4.1 and usually further purified with the help of a Wizard[™] Clean-up System (Section 2.4.4). For sequencing PCR products the DNA was purified with the help of a High Pure PCR Product Purification Kit (Section 2.4.7).

2.5.4.1 Manual sequencing

For manual sequencing the method of Sanger *et al.* (1977) was followed using the T7 Sequencing Kit (Pharmacia).

DNA denaturation and primer annealing

5-10 µg of DNA in 8 µl were mixed with 2 µl of 2 M NaOH by vortexing and left at RT for 10 min. The solution was snap-cooled on ice and 10 µl of 0.9 M sodium acetate (pH 4.8) were mixed in. The DNA was precipitated by the addition of 60 µl of ethanol and placed at -70°C for 15 min. The DNA was pelleted by centrifuging for 10 min at 14,000 x g, 4°C. The pellet was washed with ice-cold 70% ethanol and re-dissolved in 12 µl of primer solution (10 µl of H₂O + 2 µl of primer 4-5 µM). The solution was incubated sequentially at 65°C for 5 min, 37°C for 10 min, and RT for 10 min.

Sequencing reactions

For each sequencing reaction four minifuge tubes were labeled 'A', 'C', 'G' and 'T'. To each was added 2.5 µl of a supplied Termination Mix corresponding to the dideoxytermination mixes for the nucleotide indicated. 1 µl of Q H₂O was dispensed in a new microfuge tube and to it were added 3 µl of "labeling mix-dCTP", 2 µl of T7 DNA polymerase (previously diluted to 1.5 u/µl in supplied dilution buffer), and 1 µl of labeled [α -³⁵S]dATP. 6 µl of this mix were added to the annealed template and primer and the labeling reaction thus initiated was allowed to proceed for 5 min at room temperature. Meanwhile, the four Termination Mixes were incubated at 37°C and to each, 4.5 µl of the labeling reaction mix were added. The solutions were incubated at 37°C for 5 min, after which 5 µl of supplied Stop Solution were added. The terminated sequencing reactions were either immediately resolved in a sequencing gel or stored at -20°C until used.

Preparation of polyacrylamide sequencing gels

For preparation and electrophoresis of sequencing gels a Sequi-Gen[®] Nucleic Acid Sequencing Cell (Bio-Rad) with a cell size of 21x50 cm was used. Gels of 4.6%-6% acrylamide/bis-acrylamide (Accugel 40) in 1xTBE buffer and 7 M urea were prepared from 75 ml solutions:

	Gel type (Acrylamide/bis-acrylamide %)			
	4.6%	5.0%	5.3%	6.0%
Urea	31.5 g	31.5 g	31.5 g	31.5 g
5xTBE buffer	15 ml	15 ml	15 ml	15 ml
40% Acryl./bis-acryl.	9.37 ml	10.00 ml	11.25 ml	8.75 ml
Q H ₂ O	to 75 ml	to 75 ml	to 75 ml	to 75 ml

The gel solution was filtered through a 0.45 μm glass filter membrane (Whatman) and degassed for 5 min under vacuum. The sequencing cell was assembled after coating the top glass plate with Repellcote (BDH) and its base sealed according to the manufacturer's instructions with 15 ml of gel solution to which 75 μl of each of 25% ammonium persulfate and TEMED (Sigma) were added. The gel was poured after mixing the remaining 60 ml of solution with 60 μl of each of 25% ammonium persulfate and TEMED (Sigma). A shark tooth comb was inserted at the top of the gel. After polymerization the gel was mounted on the cell support and the buffer tanks filled with 1x TBE buffer. The gel was pre-run at 55 W until its temperature (monitored with a probe) reached 50-55°C. Samples were then loaded after denaturation by heating in a dry-bath at 90°C for 2 min and immediately snapping cool on ice. Electrophoresis was conducted at about 45-55 W.

At the end of electrophoresis the sequencing cell was pried open, the gel transferred to a 3 MM Whatman filter paper, covered with Saran wrap (Dow), and dried in a Gel Slab Vacuum Drier (Bio-Rad). The dry film was autoradiographed usually for 24-48 h at RT (Section 2.8.5).

2.5.4.2 Automated sequencing

For automated sequencing a PRISM™ Ready Reaction DyeDeoxy™ Terminator Cycle Sequencing kit (Perkin-Elmer) was used. Purified double-stranded plasmid DNA (1-2 μg) were mixed with 3.2 pmol of sequencing primer and Q H₂O to a total 10.5 μl . This solution was mixed with 9.5 μl of terminator pre-mix supplied with the kit (containing dNTPs, DyeDeoxy terminators and AmpliTaq DNA polymerase in a buffer). The mix was subjected to PCR in a thermal cycler (Perkin-Elmer) after overlaying with mineral oil. The temperature conditions included a 3 min denaturation period (96°C), 25 cycles of amplification (96 °C 30 sec., 50 °C 15 sec., 4 min 60°C), followed by cooling and soaking at 4°C. The mineral oil was extracted with 100 μl of chloroform after diluting the solution with 80 μl of Q H₂O. The solution was then extracted twice with 100 μl of a (68:18:14) mix of phenol:water:chloroform to remove unincorporated DyeDeoxy terminators, according with the supplied instructions. The DNA was precipitated from the aqueous phase after transfer to a new microfuge tube by the addition of 15 μl of 2 M sodium acetate (pH 4.5) and 300 μl of ethanol. To aid the precipitation the mix was incubated at -20°C for at least 30 min. The DNA was pelleted by centrifuging 14,000 x g for 15 min at 4°C. After washing and drying the sample was resolved and analyzed on an Applied Biosystems Model 373A DNA Sequencing System by the Protein and Nucleic Acid Chemistry Laboratory of Leicester University. Some of the samples were sequenced by the Sequencing Unit at the University of Durham with a similar system. In these instances the purified DNA samples were made 2 $\mu\text{g}/\mu\text{l}$ in Q H₂O or TE buffer and supplied to the Unit together with the required primer diluted to 3.2 pmol/ μl . The Unit performed all sequencing reactions and gel analysis.

2.5.5 PCR amplification

2.5.5.1 General procedures

For the majority of PCR amplification reactions Taq DNA polymerase (Promega or Bioline) was used in reaction volumes of 25 μ l-100 μ l in the buffers supplied by the manufacturers (at 1x strength). The reaction conditions included deoxynucleotide concentrations of 0.2 mM each, primer concentrations of 0.2 μ M, and 0.2-1 units of enzyme. Various MgCl concentrations were tested for each primer pair but usually 1.5 mM was found to be the most suitable.

Cycling was performed in a Perkin-Elmer DNA Thermal Cycler 480. Cycling parameters varied with the primers and experiments. In most cases a hot-start scheme was followed whereby the enzyme to be added to a reaction was diluted to a final 5 μ l in 1x buffer at the appropriate MgCl₂ concentration. The diluted enzyme was then mixed into the PCR reaction pre-heated to 85°C and cycling initiated.

2.5.5.2 Amplification of genomic DNA

For PCR amplification of specific genomic fragments of unknown size or larger than 3 kb, a system optimized for amplification of fragments up to 12 kb, Expand™ High Fidelity PCR System (Boehringer-Mannheim), was used. The system incorporates two enzymes, namely Taq and Pwo DNA polymerases. The reaction conditions were as indicated by the manufacturer but the suggested cycling parameters were modified.

Genomic DNA was diluted to about 300 ng/ μ l. The reactions were set-up in 50 μ l of an aqueous solution including 5 μ l of 10x amplification buffer (supplied), 0.2 mM of each dNTP and 0.3 μ M of each primer, 0.75 μ l of Expand™ High Fidelity PCR System enzyme mix, and containing 1 μ l of genomic DNA. Prior to cycling the template was denatured by heating for 2 min at 94°C.

Cycling conditions were 10 x (94°C for 15 secs, annealing for 30 secs, 68°C for 2 mins 30 secs), followed by 8 x (94°C for 15 secs, annealing for 30 secs, 68°C for t). Time t was defined as $t=2$ mins 30 secs + $n \times 20$ secs, where n increase from 1 to 8 with the cycle number. Finally, a third set of cycling parameters were applied: 14 x (94°C for 15 secs, annealing for 30 secs, 68°C for 6 min). Annealing temperatures varied with the primer pairs used.

2.5.5.3 Cloning of PCR products

For cloning of PCR products generated with Taq polymerase a TA-cloning® kit (Invitrogen) was used. The PCR products were directly mixed with the pCR™II vector supplied with the kit and a ligation performed overnight at 14°C with T4 DNA ligase as recommended by the manufacturer. All reagents used were supplied.

Cloning of PCR products obtained by amplification of genomic DNA with Expand™ High Fidelity PCR System (see above, Section 2.5.5.2), required their prior purification with

High Pure PCR Product Purification Kit (Section 2.4.7). These products have a mix of blunt and 3' single dA overhang ends. To ensure the presence of 3' single A overhangs at most ends the DNA solution was made 1x in Taq PCR buffer supplemented with dNTPs and MgCl_2 , after which Taq polymerase added, as for a normal PCR reaction. The reaction was incubated at 72°C for 15 min and then cloning was performed according to the general procedure.

Transformation of the ligated products into *E. coli* used competent cells of INV α F' also provided in the kit, according to the instructions. This involved cooling the ligation reaction and thawing an aliquot of frozen competent cells on ice. To the thawed cells 2 μl of 0.5 M β -mercaptoethanol were added-in followed by 1 μl of ligation reaction. The vial was incubated on ice for 30 min after which the cells were heat-shocked in a water bath at 42°C for 30 sec. The tube was quickly cooled on ice for 2 min and 450 μl of SOC medium pre-warmed to RT were added. This cell suspension was incubated with shaking at 37°C for 1 h. Then the cells were spread onto LB agar plates supplemented with ampicillin-methicillin and with IPTG and X-Gal for detection of recombinant clones. After overnight incubation clones of interest were selected, isolated and analyzed.

2.5.5.4 Colony-PCR

Colony-PCR was used to quickly confirm the presence of inserts in recombinant plasmids and estimate their size. This involved setting-up PCR reaction cocktails (20 or 45 μl) lacking Taq polymerase and mixing in a very small amount of a colony to be tested with the aid of a sterile pipette tip. The suspension was then heated to 94°C for 10 min, 5 μl of Taq polymerase diluted as described above (Section 2.5.5.1) were added and the amplification initiated. The reaction was performed under the usual conditions for the primers used for up to 25-30 cycles. Products of the reactions were examined by gel electrophoresis.

2.5.6 Assembly of c-myc fragment

The c-myc fragment for the construction of recombinant genes encoding tagged proteins was assembled from two nearly complementary primers, prMYC-F and prMYC-R (Section 7.8.3). For assembly approximately 6 μg of each primer from 0.3 $\mu\text{g}/\mu\text{l}$ stocks were mixed in 100 μl of 1 x Taq buffer (50 mM KCl, 10 mM Tris.HCl pH 9.0, 0.1% Triton[®] X-100, 1.5 mM MgCl_2). The solution was heated in a Perkin-Elmer thermal cycler in succession to 96°C for 5 min, 90°C 5 min, 86°C 5 min, 82°C 5 min, after which the machine was switched off and allowed to cool to RT. The DNA was purified by ethanol precipitation and resuspension in TE. Annealing was confirmed by gel electrophoresis (data not shown).

2.6 RNA isolation and purification

2.6.1 Reducing RNAase contamination

To remove RNAase contamination from solutions used downstream from the extraction procedure, they were treated with DEPC, or, where not possible, prepared with DEPC-treated water. Treatment involved mixing-in of 0.1% of DEPC with continuous stirring for 1 hour followed by autoclaving to eliminate the DEPC as recommended by Sambrook *et al.* (1989).

Materials used were usually of sterile disposable plastic. Otherwise they were made RNAase-free by baking.

2.6.2 Phenol/SDS method

Solutions

TLE buffer (0.2 M Tris.HCl, 0.1 M LiCl, 5 mM EDTA): the solution was made pH=8.2 with HCl.

TLE-equilibrated phenol: prepared by washing liquid phenol three times with equal volumes of TLE. The pH of phenol was determined as described (Section 2.1.3).

Grinding buffer: prepared by making TLE 1% in SDS with 10% SDS stock solution.

8 M LiCl

Preparation

1-2 g of plant material were weighed and frozen in liquid nitrogen. The material was ground to a fine powder with the help of pre-cooled mortar and pestle.

The frozen powder was transferred into a new 50 ml plastic tube containing 15 ml of grinding buffer plus 5 ml of TLE-equilibrated phenol. The mixture was homogenized with a Polytron at moderate speed for 1 min. 5 ml of chloroform were added and mixed by inversion.

The mix was incubated at 50°C in a water-bath for 20 min. The phases were separated by centrifugation at 14,000 x g for 15 min at 4°C.

The aqueous phase was transferred to a new plastic tube and extracted with 5 ml of TLE-equilibrated phenol. 5 ml of chloroform were mixed in by shaking. The mix was centrifuged at

14,000 x g for 15 min at 4°C and the aqueous phase recovered. The extractions were repeated until no interface was visible (usually three times was sufficient). A final extraction with 5 ml chloroform was performed and the aqueous phase transferred to a new tube.

1/3 vol. of 8M LiCl were added to the solution and the tube was left on ice overnight. The precipitate was pelleted by centrifugation at 14,000 x g for 20 min. at 4°C. The supernatant was removed and the pellet washed with 2 M LiCl.

The RNA pellet was re-dissolved in 0.8 ml of DEPC-H₂O and the solution was transferred to a minifuge tube. After making it 2M in LiCl by addition of 1/3 vol. of 8 M LiCl the solution was kept on ice for 6 hours. The RNA precipitate was recovered by centrifuging at 14,000 x g for 20 min. at 4°C. The supernatant was removed and the pellet was rinsed with 2 M LiCl.

The pellet was re-dissolved in 400 µl of DEPC-H₂O and precipitated by addition of 40 µl of 3M sodium acetate and 2 vol. of ethanol. The RNA was left to precipitate overnight and recovered by centrifugation for 15 min. at 14,000 x g for 30 min, 4°C. The RNA was resuspended in an appropriate volume of TE buffer and its concentration determined by spectometry.

2.6.3 PIC method

Solutions

Extraction buffer:

50 mM	Tris.HCl, pH 8.3 (from 1 M stock)
150 mM	NaCl (from 1.5 M stock)
10 mM	EDTA, pH 8.3 (from 0.5 M stock)
1%	SDS
0.7%	β-mercaptoethanol
1%	polyvinylpyrrolidone (40,000; added just before use)
PIC	mix of phenol (24): isoamyl alcohol (1): chloroform (24), pH>8.0 equilibrated with T ₁₀₀ E ₁₀ (pH 8.3) by washing at least 4 x. pH determined as described (Section 2.1.3).

T₁₀₀E₁₀

100 mM	Tris.HCl, pH=8.3 (from stock of 1M Tris.HCl)
10 mM	EDTA, pH=8.3 (from stock of 0.5 m EDTA)

T₁₀E₁ prepared by dilution (10x) of T₁₀₀E₁₀

T₁₀E₁₀

10 mM Tris.HCl, pH=8.3 (from stock of 1M Tris.HCl)

10 mM EDTA, pH=8.3 (from stock of 0.5 M EDTA)

3 M sodium acetate, pH 5.2

8 M LiCl

Preparation

This method was used for very small amounts of material. The material was kept on ice while collecting (*e.g.* while siliques or seeds were being dissected), and then frozen in liquid nitrogen, or immediately frozen when possible.

Weight was estimated and when at least 300 mg or more were available, the material was ground to a fine powder with the help of pre-cooled mortar and pestle. About 50-100 mg were transferred to a minifuge tube containing 150 µl of extraction buffer and 150 µl of PIC and immediately vortexed for 1-2 min. The phases were separated by centrifuging at 14,000 x g for 5 min at 4°C. The upper phase was collected and transferred to a new minifuge tube containing 150 µl of PIC and the extraction procedure was repeated. The process was re-iterated until no interphase was observed.

After recovery of the aqueous phase into a new tube the nucleic acids were precipitated by the mixing in 0.1 vol. of 3M sodium acetate and adding 1 vol. isopropanol. The solution was left for a few hours to overnight at -20°C.

The nucleic acids were pelleted by centrifuging at 14,000 x g for 30 min at 4°C. The pellet was washed with 80% ethanol and then completely re-dissolved in a minimal volume of T₁₀E₁₀ (up to 1125 µl). To this solution 1/3 vol. of 8 M LiCl were added and the mix was left on ice overnight.

The solution was centrifuged at 14,000 x g for 30 min at 4°C and the supernatant removed. The pellet was re-dissolved in half of the volume of T₁₀E₁₀ previously used. 1/3 vol. of 8 M LiCl was then mixed in and the solution left to stand on ice for 6 h. The solution was centrifuged at 14,000 x g for 30 min at 4°C and the pellet washed with 80% ethanol. The RNA pellet was dissolved in TE buffer and stored at -70°C until further use.

2.6.4 Isolation of poly(A)⁺ RNA

Poly(A)⁺ RNA was isolated with the help of a "Dynabeads mRNA Isolation Kit" (Dyna). The method of isolation relies on the specific annealing of poly(A)⁺ RNA to oligo(dT)₂₅ tails fixed on solid, magnetic beads (Dynabeads), under appropriate conditions. The instructions provided by the manufacturer's were followed throughout.

Essentially, about 200 µg of total RNA in aqueous solution were diluted to about 0.75 µg/µl. The solution was heated to 65°C for 2 min and then added to the 0.4 ml of Dynabeads Oligo(dT)₂₅ suspension in 200 µl of supplied 2x Binding Buffer (100 mM Tris.HCl pH 8.0, 500 mM LiCl, 10 mM EDTA, 1% LiDS, 5 mM DTT). Annealing was allowed to proceed at RT for 5 min with continuous rotation. The particles were concentrated by placing the tube in a magnetic field provided by a Dynal MPC unit, the buffer replaced with 0.4 ml of Washing Buffer provided (10 mM Tris.HCl pH 8.0, 0.15 M LiCl, 1 mM EDTA, 0.1% LiDS), and the particles resuspended. The washing step was repeated. After removal of the Washing Buffer the particles were resuspended in 25 µl of Elution Solution (2 mM EDTA pH 8.0), and the tube incubated at 65°C for 2 min. After magnetically trapping the particles the buffer was transferred to a new tube. The concentration of the solution was determined by reading the A₂₆₀ of 4 µl.

2.7 RNA manipulation and analysis

To lessen the danger of RNAase contamination materials and solutions were treated as described in Section 2.6.1.

2.7.1 Denaturing agarose gel electrophoresis

For size separation of RNA molecules, electrophoresis in denaturing formaldehyde agarose gels was used.

Solutions

10 x MOPS buffer:

- 0.2 M MOPS (pH 7.0)
- 50 mM Sodium acetate
- 10 mM EDTA (pH 8.0)

Ficoll-Dye-EDTA Loading buffer (FDE):

- 0.1 M EDTA (pH 8.0)
- 0.25% Bromophenol blue
- 0.25% Xylene cyanol FF
- 0.75 M Ficoll 400

Formamide sample buffer (FSB):

10x MOPS	100 μ l
Formamide (47% stock)	200 μ l
Formaldehyde (37% stock)	120 μ l

Gel preparation

All steps were carried out in a fumehood because of the toxic nature of the formaldehyde fumes. The gel was prepared by melting 1 g of agarose in 87 ml of DEPC-H₂O, cooling to about 65-70°C and then mixing-in 10 ml of 10x MOPS and 5.1 ml of 37% formaldehyde stock. The solution was poured onto the gel tray and allowed to cool and solidify for 1 hour.

The solidified gel was placed in the electrophoresis tank and submerged in 1x MOPS buffer. While the gel was pre-run for about 10 min the samples were prepared. Prior to sample loading the gel wells were flushed with gel running buffer.

Sample preparation and resolution

Up to 15 μ g of RNA were made up to 7 μ l in DEPC-H₂O. This was mixed with an equal volume of FSB, heated to 65°C for 10 min and snapped cool on ice. 1-3 μ l of FDE and 1 μ l of ethidium bromide solution (1 mg/ml) were added to the RNA sample. The sample was loaded and electrophoresis initiated. Gels were run for about 3-6 hours at 3-4 V/cm.

2.7.2 Northern blot preparation

Northern blots of RNA denaturing gels were prepared by capillary transfer of the resolved RNA to Zeta-Probe[®] GT membranes (Bio-Rad), essentially as described for Southern blots (Section 2.5.3) but with modifications. The gels were not pre-treated with HCl and, rather, were directly placed over the source of transfer solution. The transfer solution was 50 mM NaOH. Blotting was performed for 8-12 hours. The membrane was briefly rinsed in 2 x SSC previously treated with DEPC, air dried and baked in a vacuum oven at 80°C for 1 h. The baked membranes were stored at RT until use.

2.7.3 RT-PCR

Reverse transcription reaction

For reverse transcription SuperScript[™] II (Gibco/BRL) was used. For each reaction 1 μ g of purified total RNA and 1.0 μ l of template-specific primer (10 μ M) were diluted with sterile Q H₂O to 11 μ l (or 12 μ l for negative controls lacking enzyme). The solution was overlaid with mineral oil, heated to 70°C for 10 min in a dry bath, and quickly chilled on ice. A freshly prepared mix of 4 μ l of 5x First Strand Buffer (supplied with the enzyme; 250 mM Tris.HCl,

pH 8.3, 375 mM KCl, 15 mM MgCl₂), 2.0 µl of 0.1 M DTT, and 1.0 µl of a dNTP mix (10 mM each, at pH 7.0) was added. The solution was incubated at 48°C for 2 min, after which time 1.0 µl (200 U) of SuperScript™ reverse transcriptase were added to the reactions (but not to the negative controls), and 1 µl of RNasin® ribonuclease inhibitor (40 U/µl; Promega). The reaction was allowed to proceed at 48°C for 1 h and then stopped by heating to 70°C for 15 min.

For semi-quantitative PCR 1 µl of SuperScript™ enzyme storage buffer (20 mM Tris.HCl pH 7.5, 100 mM NaCl, 0.1 mM EDTA, 1 mM DTT, 50% glycerol) was added to the quantitation controls. These were set-up as the reverse transcription reactions except that, in addition to the RNA they contained either 10⁴, 10⁵ or 10⁶ molecules of pCEN1-8 or pCEN1-9 linearized with *EcoR* I.

PCR amplification

Amplification of cDNA fragments was performed with Taq DNA polymerase (Promega) with specific primers. In a 500 µl microfuge tube 2.0 µl of cDNA solution were mixed with 4.0 µl of 10x reaction buffer (500 mM KCl, 100 mM Tris.HCl pH 9.0, 1.0% Triton® X-100), 2.8 µl of 25 mM MgCl₂, 1.0 µl of dNTP mix (10.0 µM each), 1.0 µl of each of the primers (10.0 µM), and Q H₂O to 49.8 µl. The solution was overlaid with two drops of mineral oil and incubated at 95°C for 2-3 min in a Perkin-Elmer thermocycler. At this stage the reaction was initiated by mixing in 0.2 µl of Taq DNA polymerase (5 U/µl) and starting the cycling program. Cycling conditions varied with the primers used. When necessary, the volumes were increased maintaining the concentrations of the various reagents.

When two successive PCR amplifications of the same cDNA were performed, the second reaction was initiated as a 1/10 dilution of the first with the necessary adjustments in concentration of buffer and MgCl₂.

Where thought necessary the amplification of genomic fragments was achieved by digestion with *Acc* I for which a site is present in 3 bp downstream from exon 2 of *PKT1/PKT2*. The digestion was performed by adding 1 µl of *Acc* I (10U/µl) to the reverse transcription reaction and incubating for 1-2 hours at 37°C.

2.7.4 RACE of 5' ends of transcripts

2.7.4.1 SLIC method

Reverse transcription

About 1-2 µg of poly(A)⁺ RNA, prepared as described earlier (Section 2.6.4), were mixed with 1.0 µl of the cDNA synthesis primer (10 µM) and DEPC-treated H₂O to a final volume of 10.0 µl. A positive control reaction was set-up with a supplied human placental poly(A)⁺ RNA

sample. The primers utilized were, respectively, prEN1-P3, prGUS-cDNA, and 3'-TFR1 for the poly(A)⁺ RNA samples from *A. thaliana* cv. Columbia, AtEN101 and the control placental RNA. Replicas were prepared for negative controls of the two experimental reactions with about 0.5 µg of poly(A)⁺ RNA. The mixes were incubated at 65°C for 10-15 min. During this interval a master mix A was made consisting of:

46.0 µl	DEPC-treated Q H ₂ O
45.0 µl	4x Reverse transcriptase buffer
8.0 µl	RNAse inhibitor (40 units/µl)
18.5 µl	Ultra pure dNTP mix (10 µM)

A master mix B was prepared by adding 1.5 µl of AMV reverse transcriptase (25 units/µl) to 70.5 µl of master mix A. The master mix B was placed in a hot block at 52°C and left for at least 3 minutes to equilibrate the temperature. The tubes containing poly(A)⁺ RNA were transferred to the 52°C block for 3 minutes and then the test reactions were initiated by adding to each 20.0 µl of master mix B. To the two negative controls, 20 µl of master mix A were added. The reactions were terminated at the end of 60 minutes incubation by adding 1.0 µl of 0.5 M EDTA.

RNA hydrolysis and cDNA purification

2.0 µl of the cDNA synthesis reactions were saved and the remainder 28 µl were mixed with 2.0 µl of 6 N NaOH. The new solution was incubated at 65°C for 30 minutes to hydrolyze the RNA. After neutralizing with the addition of 2.0 µl of acetic acid 6 N the solutions were mixed with 80.0 µl of 6 M NaI.

For purification, 8.0 µl of a supplied silica-matrix suspension (GENO-BIND™) were mixed with the cDNA solution. After incubation on ice for 10 minutes with occasional mixing, the suspension was spun at 12,000 x g, 4°C in a tabletop centrifuge for 10 secs. The pellet was washed twice with 500 µl of 80% ethanol and then let air dry for 2 min. To recover the cDNA, the pellet was resuspended in 50 µl of DEPC-treated Q H₂O, incubated at 65°C for 5 min, and spun at 12,000 x g, RT for 2 min. 45 µl of the cDNA-containing supernatant were transferred to a tube containing a mix of 2.0 µl (10 mg/ml) of glycogen, 5.0 µl of 2 M sodium acetate and 100.0 µl 95% ethanol. After incubation for 30 min at -20°C the cDNA was pelleted by centrifuging at 14,000 x g, 4°C for 30 min, washed with 40.0 µl of 80 % ethanol and resuspended in 6.0 µl of DEPC-Q H₂O.

Anchor ligation

To add the AmpliFINDER Anchor to the 3' end of the cDNAs the following components were mixed together per reaction:

2.5 µl	cDNA solution
2.0 µl	AmpliFINDER Anchor (2 µM)
5.0 µl	2x single-stranded ligation buffer

Ligations in the test and positive control samples were carried out by addition of 0.5 µl of T4 RNA ligase (20 units/µl). 0.5 µl of sterile Q H₂O were added to the two negative controls. For the *PKT1/PKT2* and the control cDNAs the solution was incubated at room temperature for 20 hours and then diluted 1:10 with sterile Q H₂O, as recommended by the manufacturer of the kit. Ligation of the Anchor to the fusion cDNA was carried out for 48 hours with an additional 10 units of T4 RNA ligase being added at the end of 24 hours.

PCR amplification

PCR reactions were in 1x PCR buffer (50 mM KCl, 10 mM Tris.HCl pH=9.0, 0.1% Triton[®] X-100), 1.5 mM MgCl₂, 0.2 mM dNTPs, 0.2 µM in each primer and 2.5 units of Taq polymerase (Promega). Reactions were 50 µl in volume including 1.0 µl of cDNA solution or negative control, or 0.5 µl for the human cDNA control. Reactions were initiated with hot-start, *i.e.* the Taq polymerase was added after the reaction mixes had been equilibrated to 85°C. Where two sequential rounds of amplification were used, 5.0 µl of the first reaction were used to seed the second amplification.

PCR products were cloned following ligation with pCR[™]II vector and transformation of DH5α cells.

***PKT1/PKT2* cDNAs**

Oligonucleotide primers used:

prEN1-P2
prEN1-P1

Cycling conditions:

Round 1: 1 x 94°C 2 min
 36 x (94°C 45 secs, 60°C 45 secs, 72°C 2 min)
 1 x 72°C 10 min

Round 2: 1 x 94°C 2 min
 36 x (94°C 45 secs, 60°C 45 secs, 72°C 2 min)
 1 x 72°C 10 min

Fusion transcripts

To confirm cDNA synthesis a PCR amplification was performed with primers prGUS-cDNA and prGUS-CONTROL. These primers were used to amplify 1.0 µl aliquots of the reverse transcription reactions in 50 µl. The cycling parameters were: 38 x (94°C 45 secs, 66°C 35 secs, 72°C 1 min).

Amplification of SLIC products of fusion transcripts involved two rounds of amplification including:

Round 1: with prGUS-cDNA and AmpliFINDER primer

1 x 94°C 2 min
 40 x (94°C 45 secs, 64°C 45 secs, 72°C 2 min)
 1 x 72°C 10 min

Round 2: with prGUS-PCR and AmpliFINDER primer

1 x 94°C 2 min
 38 x (94°C 45 secs, 66°C 45 secs, 72°C 2 min)
 1 x 72°C 10 min

For cloning the three first PCR reactions were set-up and cycling was as above but for a total of 37 cycles. The second round of amplification was performed in 15 parallel reactions as detailed above but lasting 37 cycles. The products of all reactions were pooled together, purified and concentrated by ethanol precipitation prior to cloning.

Positive control with human transferrin cDNA

To check the efficiency of the various steps in the 5' RACE procedure with the human transferrin cDNA the recommended procedures were followed. The combinations of primers used were:

Test

cDNA synthesis

cDNA purification

Anchor ligation

Presence of cDNA in Anchor ligation reaction

Primers

TFR2 and TFRQC

TFR2 and TFRQC

TFR2 and AmpliFINDER Anchor primer

TFR2 and TFRQC

Each PCR reaction was performed with 0.5 µl samples taken at the steps to be tested. Amplification was executed in a single round of 25 x (94°C 45 secs, 60°C 45 secs, 72°C 2 min) for the reactions using the primer pair TFR2 and TFRQC. The same conditions were applied for the test for Anchor ligation but the amplification lasted 35 cycles.

2.7.4.2 LOPRACE

This method consisted of three stages: A 1st strand cDNA synthesis by reverse transcription is followed by 2nd strand cDNA synthesis with the a modified Klenow fragment of DNA polymerase, Exo⁻ Klenow (Stratagene). Priming of 2nd strand synthesis is provided by looping with hairpin formation at the 3' end of the 1st strand cDNA. Effectively, this results in the formation of a single-stranded cDNA molecule with dyad symmetry. The third stage consists in the amplification by PCR of cDNA molecules, which requires a single primer.

1st strand cDNA synthesis

Total RNA of *A. thaliana* cv. Columbia was isolated from 10 day-old seedlings grown under continuous light in 1/2 MS₁₀ as described in Section 2.1.4. For first strand cDNA synthesis, the instructions suggested by the reverse transcriptase manufacturer were followed with slight modifications. Reverse transcription was performed with SuperScript™ II (Gibco/BRL), a modified, cloned Moloney Murine Leukemia Virus reverse transcriptase. A solution of 2.0 µg of total RNA and 4 pmol of prEN1-P3 in sterile DEPC-Q H₂O was prepared. This solution was equally divided into two 0.5 ml tubes. Both solutions were incubated at 70°C for 10 minutes. During this period Solution A was prepared.

Solution A (per reaction):

- 4.0 µl 5 x Reverse transcription buffer
- 2.0 µl DTT
- 1.0 µl dNTPs mix (10 mM)

At the end of the incubation period, the RNA solutions were immediately cooled on ice for about 2 minutes during which time 7.0 µl of mix A were added to each tube. The tubes were then incubated at 45°C for 2 minutes to equilibrate the temperature of the solutions. To one of the tubes 1.0 µl of RNasin® (40 units/µl; Promega) followed by 1.0 µl of SuperScript™ II reverse transcriptase (200 units/µl) were added. To the second tube, a negative control, the final volume of 20.0 µl was completed with sterile DEPC-treated Q H₂O. Both solutions were layered with a drop of mineral oil. The incubation at 45°C was continued for 1 hour.

2nd Strand cDNA synthesis

At the end of the first strand synthesis, 9.5 µl of each tube were mixed with 9.5 µl of solution B (50 mM Tris.HCl pH=6.7, 62 mM KCl, 20 mM MgCl₂) in a new tube. The new mixes were incubated at 96°C for 3 minutes to inactivate the reverse transcriptase and disrupt the secondary structure of the cDNA. At the end of the incubation period the solutions were immediately cooled on ice and to each 0.5 µl of a dNTP mix (10 mM) and 0.5 µl of Exo⁻ Klenow (5 U/µl) were added in. The tubes were transferred to a Perkin-Elmer thermal cycler and incubated sequentially at: 14°C 10 min, 15°C 2 h, 17°C 30 min, 20°C 1 h, and 70°C 10 min.

PCR amplification

For the first amplification of the cDNA sample and the negative control, 2.5 µl aliquots of each were added to 47.5 µl of PCR reaction solution. The final solution was 1 x PCR buffer (Promega), 1.5 mM MgCl₂, 0.2 mM dNTPs and 0.2 µM in primer prEN1-P2. Cycling was for 32 cycles (94°C 45 secs, 60°C 30 secs, 72°C 90 secs). The second amplification reactions were seeded with 2.5 µl aliquots of the first amplification reactions and performed under identical conditions. Cycling lasted for 28 cycles.

2.7.4.3 5' RACE with homopolymeric tailing

For implementation of this method a "5' RACE System for Rapid Amplification of cDNA Ends" kit from Gibco/BRL (Cat. No. 18374-025) was used. The various procedures followed the instructions provided with the kit with some modifications.

cDNA synthesis

Total RNA from wildtype *A. thaliana* cv. Columbia and from AtEN101 8 or 10 day-old seedlings grown in 1/2 MS₁₀ under continuous light was used (Section 2.6.2). Synthesis of first strand cDNA was essentially identical to that described above for LOPRACE. However, for reverse transcription SuperScript™ II was used (200 U/µl) and at the end of the incubation at 45°C the reactions were heated at 70°C for 15 min to stop the reaction. To hydrolyze the RNA, the reaction tubes were equilibrated to 55°C for 2 min. Then, 1.0 µl of RNase H (2 U/µl) were mixed in and the incubation continued for 10 min.

Isolation and purification of the cDNA

To the cDNA solutions, 120.0 µl of binding solution (NaI 6 M) were added. The mix was transferred to a GlassMAX spin cartridge supplied with the kit. The cartridge was spun at 13,500 x g for 30 sec. The GlassMAX silica-based matrix was washed three times with 400 µl of a supplied cold (4°C) 1 x Wash Buffer, and twice with 400 µl of cold (4°C) 70% ethanol. After centrifuging the cartridges for 1 min at 13,500 x g to remove the residual ethanol, the bound cDNA was eluted into pre-heated (65°C) 50.0 µl of sterilized Q H₂O.

Tailing of cDNA with terminal deoxynucleotidyl transferase (TdT)

A tail of homopolymeric dC was added to the 3' end of the cDNAs with TdT by following the recommended procedures. For tailing the following components were mixed together:

6.5 µl	DEPC-treated H ₂ O
5.0 µl	5 x tailing buffer (50 mM Tris.HCl pH=8.4, 125 mM KCl, 5 mM MgCl ₂)
2.5 µl	dCTP 2 mM
10.0 µl	cDNA solution

The mixes were covered with a drop of mineral oil, and incubated at 94°C for 2 min. The solutions were quickly chilled on ice for 3 min and then 1.0 µl of TdT (10 units/µl) was gently mixed in. The tubes were transferred to 37°C and incubated for 10 min.

To inactivate the TdT, the tubes were placed at 70°C for 10 min and then cooled on ice.

PCR amplification of dC-tailed cDNA

For the first PCR amplification of all dC-tailed cDNA and control samples, 5.0 µl of each were mixed with:

29.0 µl	sterile Q H ₂ O
5.0 µl	10 x PCR buffer (500 mM KCl, 100 mM Tris.HCl pH=9.0, 1% Triton [®] X-100)
3.0 µl	MgCl ₂ (25 mM)
1.0 µl	10 mM dNTP mix
2.0 µl	anchor primer (10 µM)
1.0 µl	cDNA-specific primer

After overlaying with mineral oil the reaction mixes were incubated at 94°C for 5 min and then placed at 85°C. Meanwhile, *Taq* DNA polymerase was diluted to 0.4 U/µl in 1 x PCR buffer and 5.0 µl were added to the reaction mixes and the amplification initiated.

Fusion cDNA

The first PCR amplification was carried out for 32 cycles (94°C 35 secs, 66°C 30 secs, 72°C 2 min) with the cDNA-specific primer prGUS-PCR.

A second set of PCR reactions were performed. These reactions were seeded with 5.0 µl of the first reactions and made up to 50.0 µl of 1X PCR buffer (50 mM KCl, 10 mM Tris.HCl pH=9.0, 0.1% Triton[®] X-100), 1.5 mM MgCl₂, 0.2 mM dNTPs, 0.2 µM in LB-267, 0.4 µM in anchor primer and 2.5 units of *Taq* polymerase (Promega). Amplification was for 25 cycles

(94°C 35 secs, 60°C 30 secs, 72°C 2 min). Gel resolution of the reaction products revealed abundant fragments of 0.2 kb and 0.3 kb in addition to a faint band of 0.7 kb fragments (Fig. 6.4A).

To obtain sufficient amounts of the 0.7 kb fragments for cloning a third PCR amplification reaction was prepared with the LB-267 primer. The gel in Figure 6.4A was blot dried and the band of 0.7 kb fragments was stabbed with a sterile pippete tip. The small portion of gel retained in the tip was transferred to 49.0 µl of PCR reaction solution kept at 95°C. The incubation was prolonged for 5 min. To the reaction, 1.0 µl of *Taq* DNA polymerase (2.5 units/µl) was added and then cycling was initiated. A total of 32 cycles were conducted (94°C 35 secs, 60°C 30 secs, 72°C 2 min).

cDNAs of *PKT1/PKT2*

For PCR amplification the primers prEN1-P3 and prTHIO -5R were utilized together with the anchor primer. A first amplification was carried out for a total of 32 cycles (94°C 35 secs, annealing 30 secs, 72°C 2 min). The annealing temperatures were 54°C with prTHIO-5R and 66°C with prEN1-P3. To eliminate competition from genomic DNA, at the end of the amplification 1 µl of *Acc* I (10 U/µl) was added and the solution incubated at 37°C for 1 h. Subsequently, 5.0 µl of each reaction were used to seed a second set of PCR amplifications. These were performed with similar parameters to the first and lasted for 20 cycles. To obtain sufficient material for cloning multiple PCR amplifications were set-up as described. The products were pooled and concentrated by precipitating.

2.8 Nucleic acid hybridization

2.8.1 Labeling of probes

For screening of libraries, and for Southern and Northern blot hybridization radiolabeled DNA probes were prepared by the random primer method (Feinberg & Vogelstein, 1984). The DNA templates used were derived from digested plasmids, resolved by agarose gel electrophoresis. The templates were purified as described below Section 2.8.2. Alternatively, when the digested plasmids were resolved in low melting point agarose gels, the bands of interest were excised in minimal agarose plugs and mixed with 3:1 (v/w) of Q H₂O. The gel was melted by heating to 65°C for 5-10 min and stored frozen until use.

For probe synthesis a Prime-It[®] II Random Primer Labeling Kit (Stratagene) was used. The reactions were conducted according to the protocol provided by the manufacturer. 25 ng purified DNA template (in TE buffer or H₂O) were used per labeling reaction. The template solution was adjusted to 23 µl with Q H₂O and mixed with 10 µl of the supplied solution of oligonucleotide primers. The DNA was denatured by boiling in a water bath for 5 min, the

condensate collected by brief centrifugation, and 10 μ l of supplied 5x primer buffer (for [α - 32 P]dCTP labeling) mixed in. 5 μ l of label [α - 32 P]dCTP (3,000 Ci/mmol) were added and the reaction initiated with 1 μ l of Exo(-) Klenow enzyme (5 U/ μ l). The reaction was allowed to proceed for about 15 min at 37°C and then stopped with the addition of 2 μ l of Stop Mix.

To prepare probes from DNA fragments isolated from low melting point agarose gels, prior to initiation of the labeling reaction the DNA solution was warmed at 65°C for 5 min. The labeling was performed as described above but conducted at 37°C for 30 min.

2.8.2 Probe purification

Unincorporated nucleotides were removed by gel filtration through a Nuc-Trap[®] Probe Purification Column (Stratagene). This involved pre-conditioning the column with 70 μ l of STE buffer (100 mM NaCl; 20 mM Tris.HCl, pH 7.5; 10 mM EDTA), loading the labeling reaction solution and forcing through another 70 μ l of STE. The eluate containing the labeled DNA fragments was collected and used in the hybridizations as described below.

2.8.3 Southern blot hybridization

This procedure applied to both Southern blots prepared as described in Section 2.5.3 as to filters of λ plaque lifts prepared during library screening (Section 2.3.2).

Solutions

Hybridization solution (0.5 M Na₂HPO₄ pH 7.2, 7% SDS)

Wash solution I (40 mM Na₂HPO₄ pH 7.2, 5% SDS)

Wash solution II (Na₂HPO₄ pH 7.2, 1% SDS)

Hybridization

Hybridizations were performed in rotating Hybaid incubators. The Southern blot membranes were placed in Hybaid bottles and pre-hybridized in about 20 ml of hybridization solution for 15-30 min at the hybridization temperature. The solution was removed and a pre-warmed fresh minimum volume (10-15 ml) added and allowed to equilibrate to the hybridization temperature. Meanwhile, the radioactively labeled probe (Section 2.8.2) was denatured by boiling for 5 min, snapped cool on ice and immediately added to the hybridization solution. Hybridization was allowed to proceed for 14-18 h. Hybridizations were mostly conducted at high-stringency with temperatures of 65°C for genomic DNA and library screens. 68°C were often used with other DNA sources. Low-stringency hybridization of genomic DNA was performed at 55 °C (Section 10.3.1).

Washing

At the end of the hybridization period, the filters were washed to remove unspecifically bound probe. All washes were conducted at the hybridization temperatures. For washing the filters were incubated in 100-150 ml of pre-warmed Wash solution I for 30-60 min. This wash was repeated once. For Southern blots of genomic DNA or for library screens a third wash was performed with pre-warmed Wash solution II for 10-15 min. In the case of the low-stringency hybridization of genomic DNA, the third wash was for 30 min in Wash solution I. Other Southern blots were washed more extensively with Wash solution II for 30-60 min, twice.

The washed filters were blotted dry and wrapped in Saran wrap (Dow) and placed in X-ray film cassettes for autoradiography.

2.8.4 Northern blot hybridization**Solutions**

20 x SSPE stock (3.6 M NaCl, 0.2 M Na_2HPO_4 pH=7.7, 0.02 M EDTA)

50 x Denhardt's reagent (1% Ficoll Type 400, 1% polyvinylpyrrolidone, 1% BSA Fraction V)

Pre-hybridization solution (50% formamide, 5 x Denhardt's reagent, 0.1% SDS, 0.01% poly(A), 5 x SSPE, 0.01% denatured sonicated herring sperm DNA). The herring sperm DNA denatured and added just before use.

Hybridization solution (50% formamide, 2 x Denhardt's reagent, 0.1% SDS, 0.01% poly(A), 5 x SSPE, 0.02% denatured sonicated herring sperm DNA). The herring sperm DNA denatured and added just before use.

Wash solution I (2 x SSC, 0.1% SDS)

Wash solution II (1 x SSC, 0.1% SDS)

Wash solution III (0.1 x SSC, 0.1% SDS)

Hybridization

Northern blots were pre-hybridized in 50 ml of pre-hybridization solution in a Hybaid rotary incubator for 4 hours at 42°C or 48°C. Prior to pre-hybridization the solution was warmed to temperature and the herring sperm DNA added after being denatured by boiling for 5 min and snapped cool on ice.

At the end of the pre-hybridization period the solution was replaced with 10 ml of pre-warmed hybridization solution. Freshly denatured herring sperm was added and the temperature allowed to equilibrate to 42°C (or 48°C). Then the previously prepared radioactively-labeled probe (Sections 2.8.1 and 2.8.2) was denatured by boiling, snapped cool on ice, and added to the hybridization solution. Hybridizations were continued at 42°C (or 48°C) for 45-48 hours after which the membranes were washed.

Washing

At the end of the hybridization period, the solution containing the probe was removed and quickly replaced with 100 ml of Wash solution I. The wash was incubated for 10 min at RT and then repeated with a fresh volume of Wash solution I. Two washes were then performed with Wash solution II under identical conditions. The membranes were finally washed twice with Wash solution III for 10 min at 48°C.

At the end of the washes the membranes were blot dried, wrapped in Saran wrap and placed in an X-ray film cassette in preparation for autoradiography (see below).

2.8.5 Autoradiography and phosphoimaging

For autoradiography Amersham Hyperfilm-MP was used as X-ray film. Blots hybridized with [α - ^{32}P]dCTP-labelled probes were in most cases exposed for 1 day to several weeks in cassettes fitted with intensifying screens and kept at -70°C. In some instances exposure was carried out for a few hours at RT, depending on the intensity of the signal.

Dry sequencing gels resulting from manual sequencing (Section 2.5.4.1) were exposed to Amersham Hyperfilm-MP at room temperature usually for 24-48 h.

In some instances detection of hybridization was performed by phosphoimaging using a GS-525 Molecular Imager System (Bio-Rad).

2.9 Biolistics for functional assays of promoter activity

The transcription promoting potentials of genomic segments linked to the reporter gene *gus* in appropriate plasmid constructs were evaluated in transient expression assays (Chapter 8).

Delivery of the constructs to the living plant material was performed by biolistics.

Materials

Particle gun: BIOLISTIC[®] PDS-1000/He (Bio-Rad)

Microcarriers: 1.6 μm Gold (Bio-Rad #1652264)

Macrocarriers: (Bio-Rad #165-2334)

Rupture disks: 1350 psi and 1550 psi (Bio-Rad)

Method

Preparation of microcarriers

60 mg of gold microcarriers (Bio-Rad) were washed and sterilized by vortexing vigorously in 1 ml of 100% ethanol. After pelleting by centrifugation the microcarriers were resuspended in fresh 100% ethanol and kept at RT until use.

Prior to testing of a plasmid the microcarriers in the ethanol suspension were first sonicated thrice for 30 secs, centrifuged at 10,000 x g for 1 min at RT and the supernatant removed. The microcarriers were then washed by adding 1 ml of sterile Q H₂O, vigorous vortexing and pelleting by centrifugation. The microcarriers were washed once again with Q H₂O and finally resuspended in 1 ml of sterile Q H₂O. While vortexing it, the suspension was aliquoted into 50 µl portions. The aliquots were either immediately utilized or stored at 4°C.

To coat the microcarriers with the DNA an aliquot was vigorously vortexed while 5 µl of plasmid solution (1 µg/µl) were added to it, followed by 50 µl of 2.5 M CaCl₂ and 20 µl of 0.1 M spermidine (free base). Vortexing was continued for 3 min and then the microcarriers were pelleted by centrifugation at 10,000 x g for 10 secs. The supernatant was removed and the pellet washed with 100% ethanol by briefly vortexing, centrifuging and removal of the supernatant. The microcarriers were then resuspended in 60 µl of 100% ethanol.

Preparation of target material

Preparation of the material was done under sterile conditions in a flowhood. For most assays leaves of young tobacco seedlings grown in MS₃₀ were used. The leaves were cut just prior to shooting and placed, with their adaxial surfaces down, on Petri dishes containing solid 1/2MS. Some assays were performed with developing seeds of *Phaseolus vulgaris*. These seeds were dissected open in two halves. The seeds were then laid with their cut surfaces up on 1/2MS agar plates. All plates containing the living target material were kept closed except during the shooting.

Loading of the particle gun and shooting

All accessories and the interior of the particle gun were wiped clean with 70% ethanol. Macrocarriers and rupture disks were washed and sterilized in 100% ethanol and left to air dry. Macrocarriers were placed in their holders prior to application of about 10 µl of a microcarrier suspension. The latter was vigorously vortexed or sonicated for 30 secs before immediate loading of the samples.

The macrocarriers and rupture disks were mounted in the particle gun and then the target material was placed in sample holder and uncovered. The chamber was closed and

evacuated to 28 mm Hg. Firing was performed by opening the valve connecting the chamber to the helium source.

After each bombardment the Petri dishes containing the targets were covered, removed from the chamber and sealed. To allow time for expression of the reporter gene, the bombarded specimens were incubated in the Petri dish at 25 °C on a day/night cycle for 18 to 24 hours. Then GUS activity was monitored histochemically as described below (Section 2.10.2).

2.10 Detection of GUS enzyme activity

2.10.1 Fluorometric detection

For quantitation of GUS activity in different organs its conversion of 4-methylumbelliferyl- β -D-glucuronide (MUG, Sigma) to the fluorogenic 4-methylumbelliferone (MU; Jefferson *et al.*, 1987) was monitored kinetically in a luminescence spectrometer equipped with a microtiter plate reader (Breyne *et al.*, 1993).

Reagents and solutions

4-methylumbelliferyl- β -D-glucuronide (MUG) 5 mM

4-methylumbelliferone (MU)

GUS buffer (50 mM NaH₂PO₄/Na₂HPO₄ pH 7.0, 10 mM β -mercaptoethanol, 10 mM EDTA 0.1% Triton X-100).

Bradford Protein Assay kit (Bio-Rad)

Bovine serum albumin, Fraction V (BSA)

GUS fluorimetry

Plant tissue samples (100-300 mg) were ground in 200 μ l of GUS buffer in an minifuge tube. The homogenate was cleared by centrifuging at 14,000 x g for 10 min, RT. 20 μ l of supernatant were added to 140 μ l of GUS buffer in microtiter plate well. The plate was placed in the microtiter reader and the solutions incubated at 37°C for 5 min. The reaction was initiated by adding 40 μ l of 5 mM MUG solution. Fluorescence was measured at t=0 and various time points with intervals that were adjusted to the level of activity in the sample, but typically were 10 or 20 min. The excitation λ was 355 nm and the emission λ was 460 nm.

For standardization parallel measurements were performed with known amounts of 4-methylumbelliferone (MU) in similar buffer conditions.

Protein assays

Protein assays were conducted with the Bradford Protein Assay Kit following the supplied instructions. To quantify the amount of protein in the samples a standard curve for protein concentration was obtained for a dilution series of BSA in Q H₂O. For the plant samples 20 µl of supernatant (from above) were well mixed with 1.6 ml of Q H₂O in a 2 ml cuvette. 400 µl of Bradfords reagent were added and the solution incubated at RT for 5 min. The absorbance at 595 nm was determined in a spectrophotometer.

2.10.2 Histochemical detection of GUS activity

To localize the activity of GUS in different organs and tissues a histochemical approach according to the method of Jefferson *et al.* (1987).

Reagents and solutions

X-gluc (5-bromo-4-chloro-3-indolyl β-D-glucuronic acid) 20 mM stock. Prepared in N,N-dimethylformamide and stored frozen at -20°C.

Reaction Buffer:

NaH ₂ PO ₄	100 mM
EDTA	10 mM
Triton X-100	0.1% (v/v)
pH=7.0	

Method

Prior to histochemical detection of GUS activity, where necessary to facilitate penetration of the reagents into the tissues, the materials were dissected (*e.g.* siliques and seeds). The living materials were then immersed in Reaction Buffer made 1 mM in X-gluc by addition of the reagent stock solution. The suspensions were vacuum infiltrated for 5 min, transferred to 37°C and left to incubate in the dark for up to 48 h. At the end of the incubation period, the materials were cleared by repeated washing in 70% ethanol, or by passing through an ethanol series up to 95%.

Photographic recording

For observation and photography of the stained materials a Zeiss Axioskop Stereo Microscope and a Olympus SZH10 Zoom Stereo Microscope were used. Photographic recording was performed with Ektachrome 160 tungsten-balanced film.

2.11 Computer-based searches, analysis and modelling

A variety of software and on-line facilities were used to retrieve and analyze data, including:

- For retrieval of sequence data GeneBank (Benson *et al.*, 1997) was used via the National Institutes of Health (USA) Entrez server (<http://www.ncbi.nlm.nih.gov>) and the *Arabidopsis thaliana* Database (AtDB; Flanders *et al.*, 1998).
- For the identification of GeneBank (Benson *et al.*, 1997) sequences similar to a probe sequence FASTA (Pearson & Lipman, 1988) and BLAST (Altschul *et al.*, 1990) searches were performed on-line at the National Institutes of Health (USA; <http://www.ncbi.nlm.nih.gov/BLAST/>) or AtDB servers (Flanders *et al.*, 1998).
- For sequence analysis including comparative studies and matches to specific sequence elements, a variety of packages were used including SeqAid™, GeneJockey™, and the GCG package (Devereaux *et al.*, 1984; Genetics Computer Group, 1994), GeneDoc (Nicholas *et al.*, 1997) and MACAW (Schuler *et al.*, 1991). Search for matches to blocks of amino acid residues in known protein sequences used the Swiss-Prot and TrEMBL databases (Bairoch & Apweiler, 1998).
- Detection of potential exons was performed with two neural network tools, NetPlantGene V2.1 (Hebsgaard *et al.*, 1996), and Grail 2 (Xue *et al.*, 1994).
- Protein secondary structure models were obtained on-line at the Institut de Biologie et Chimie des Protéines (France) by a variety of methods including an improved GOR statistical method (Garnier *et al.*, 1996; Gibrat *et al.*, 1987), the homology method of Levin *et al.* (1986), the double prediction method (Deleage & Roux, 1987a; 1987b), the statistical method for non-local interactions PREDATOR (Frishman & Argos, 1996), a neural-network method for multiple-aligned sequences, SOPMA (Geourjon & Deleage, 1994; 1995), and the PHD neural-network method (Rost & Sander, 1993, 1994; Rost *et al.*, 1994).
- For 3D protein modelling the homology-threading modelling procedure ProMod (Peitsch 1993, 1995; Peitsch & Jungeneel, 1996) was followed. Templates for modelling were the crystallographic structures of the *S. cerevisiae* peroxisomal 3-ketoacyl-CoA thiolase (Mathieu *et al.*, 1994; 1997). An initial model was built by the ProMod server based on an automatically-constructed alignment of the template and submitted sequences. For model refinement the alignment was manually optimized with the help of Swiss-PdbViewer (Guex & Peitsch, 1997) software, and submitted to ProMod. Models of dimers were built with ProMod, based on the same *S. cerevisiae* templates and using the monomer models.

- Several on-line search services were employed to detect matches to plant transcription factor binding sites including MatInspector (Quandt *et al.*, 1995), SignalScan (Prestridge, 1991) and PLACE (Higo *et al.*, 1998).

Chapter 3

Screening and phenotypic characterization of mutants

3.1 Introduction

This thesis is concerned with research performed to explore the value of the T-DNA interposon pΔ*gus*Bin19 as a promoter trap and gene tagging element, and as an insertional mutagen in the study of embryogenesis in *A. thaliana* (Section 1.4.2). Two lines of work were followed, including the characterization of a transgenic line, AtEN101, where temporally-regulated *gus* reporter gene expression had been observed during embryogeny (Topping *et al.*, 1994). In parallel, a number of additional transgenic *Arabidopsis* lines transformed with pΔ*gus*Bin19 (Clarke *et al.*, 1992) were screened for mutations affecting seedling morphology that might reflect earlier defects in embryo formation.

The present chapter reports a detailed phenotypic characterization of the transgenic line AtEN101 with particular reference to GUS activity (Section 3.2). AtEN101 was, among 430 independent primary transformants (T1) of the promoter trap vector pΔ*gus*Bin19, one of the estimated 3-4% with reporter activity in developing embryos as shown histochemically by Topping *et al.* (1994). In this line reporter activity was observed both in developing embryos and embryo sacs during a limited time window in embryogenesis. The temporal regulation of reporter expression during embryogeny allied with the observation that a single T-DNA locus was present in its genome (Topping *et al.*, 1994) made it of interest to characterize AtEN101 at the molecular level.

In addition, the identification and preliminary study of a new shape mutant of *Arabidopsis* isolated from the library of pΔ*gus*Bin19-transformed lines following screening of seedlings is described in Section 3.3. The defect in this mutant, designated as *bashful*, categorized it as a dwarf and was shown to be linked to a T-DNA insert by segregation analysis.

3.2 Line AtEN101

3.2.1 Previous studies

Histochemical detection of GUS in AtEN101 had shown that in siliques the enzyme was present in embryos and embryo sacs from the early globular to late globular/early heart stage but was absent from the suspensor and silique walls (Topping *et al.*, 1994). Fluorimetric and histochemical detection also revealed low levels of reporter activity in germinating seedlings including cotyledons, shoot apex, and, in roots in the elongation zone and root cap, but not in

the meristem. In mature plants activity had been found to be restricted to the tapetum, being absent from other organs including roots, leaves and stems.

The patterns of reporter gene expression may change in a sequence of generations of the same line (*e.g.* due to gene silencing) or as a response to altered physiological conditions. To investigate possible differences in the pattern of GUS expression in sequential generations of AtEN101 plants, histochemical and fluorimetric studies were conducted.

3.2.2 Fluorimetric studies

Fluorimetric analysis of GUS activity was conducted on several organs as described in Section 2.9.1. Enzyme activity and protein content were determined in two or three aliquots of homogenate for each sample. Four to seven samples were analyzed for each organ from a same or different plants. Organs analyzed included roots and leaves of 2-4 week-old seedlings grown on 1/2MS₁₀ as well as stems, ovules and embryos at different developmental stages, and flowers from mature soil-grown plants. Fluorimetric activity of organs from C24 plants analyzed in parallel were very low ($<2 \text{ pmol.mg}^{-1}\text{min}^{-1}$) and are not shown. The results are presented as a chart in Figure 3.1. The highest levels of activity as measured by the rate of accumulation of MU were observed for flowers and siliques with globular stage embryos (Fig. 3.1). The lowest levels were found in seeds ($6.3 \pm 1.8 \text{ pmol.mg}^{-1}\text{min}^{-1}$). Siliques with heart to torpedo stage embryos, and embryos at late cotyledonary/maturation stages showed intermediate levels of enzyme activity, which was higher in early cotyledonary embryos ($132.5 \pm 14.53 \text{ pmol.mg}^{-1}\text{min}^{-1}$)

3.2.3 Histochemical characterization

Homozygous EN101 plants of several generations including T4, T6, T8 and T10 were subjected to histochemical analysis. For all, young seedlings germinated in 1/2 MS₁₀ stained strongly in roots except the meristem, and in the shoot apex (Figs. 3.2A-F). GUS activity was observed at low level in the hypocotyl where it was concentrated in the vascular tissue, being higher in cotyledons and leaves (Fig. 3.2A and B). In older leaves the staining often was not uniform but rather patchy, tending to be retained longer in peripheral regions (Fig. 3.2E). When T8 and T10 seeds were germinated in 1/2MS medium (thus lacking the sucrose present in 1/2MS₁₀) the expression of GUS had a similar pattern to those germinated in 1/2MS₁₀ but the levels were lower as demonstrated by parallel growth of siblings seeds and histochemical processing of the germinated plants (Fig. 3.2F). Germination in the dark resulted in normal skotomorphogenic development with etiolation, elongation of the hypocotyl, absence of greening and of opening of the cotyledons. GUS activity was observed but limited to the cotyledons, irrespective of the presence of sucrose in the media (Fig. 3.2G).

In mature plants grown in soil GUS activity was found in mature and senescent leaves, again with a non-uniform distribution and tending to persist longer in marginal areas (Fig.

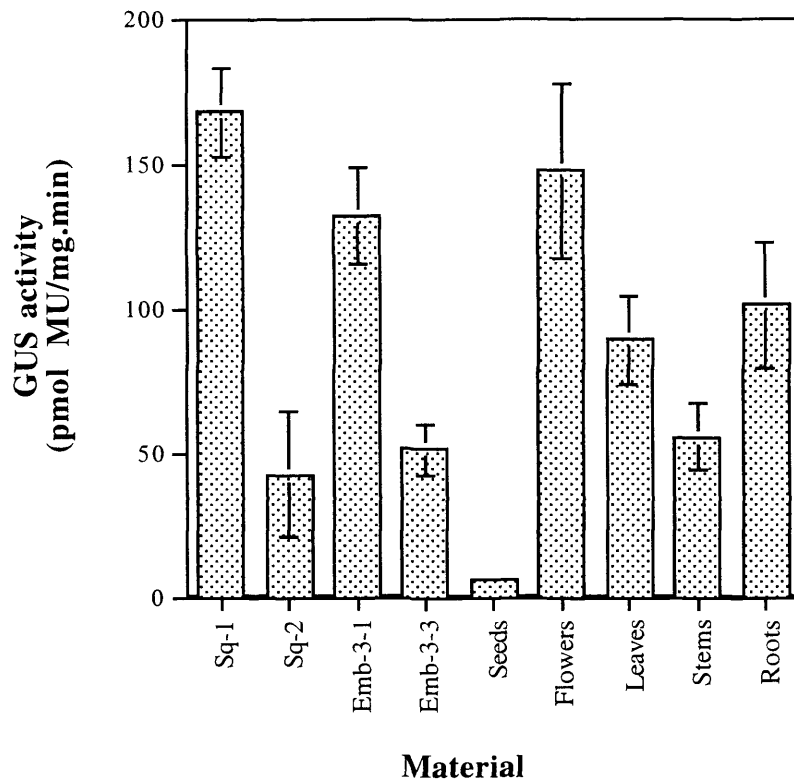


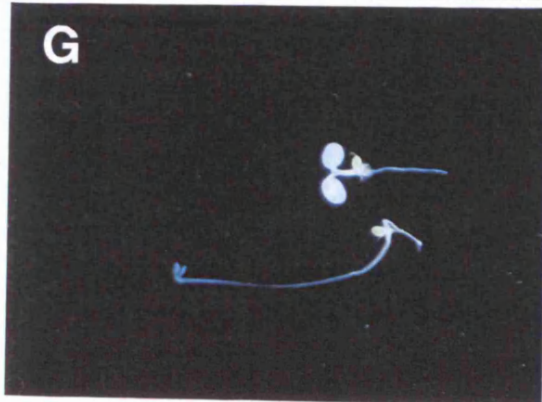
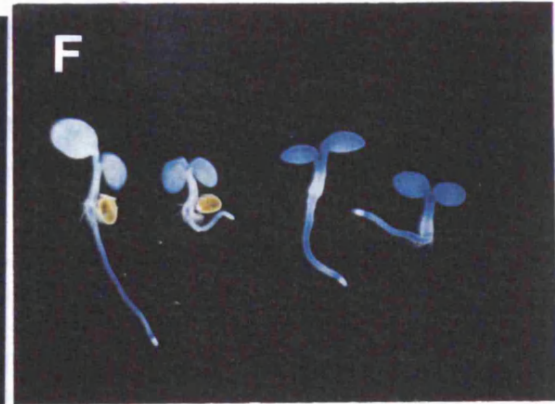
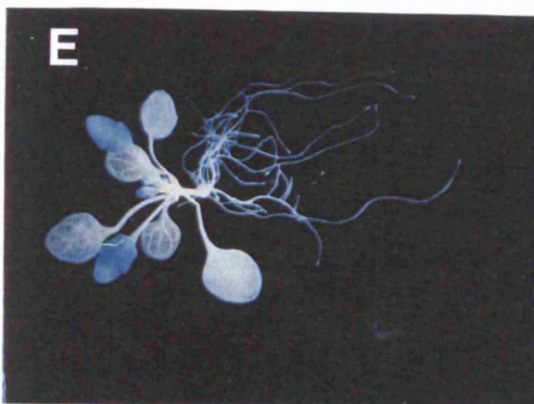
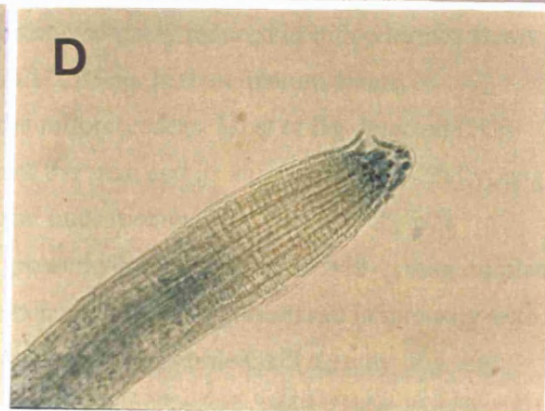
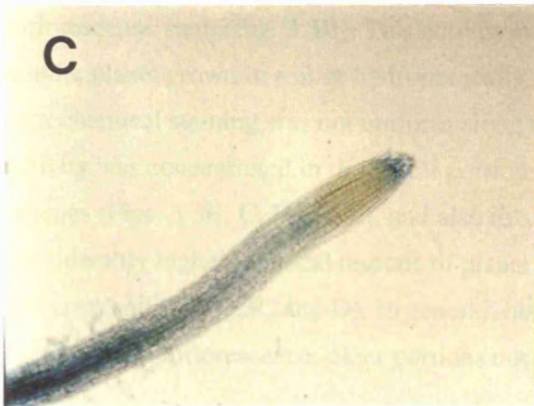
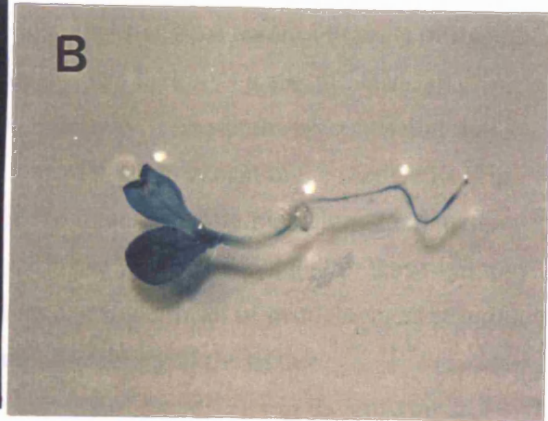
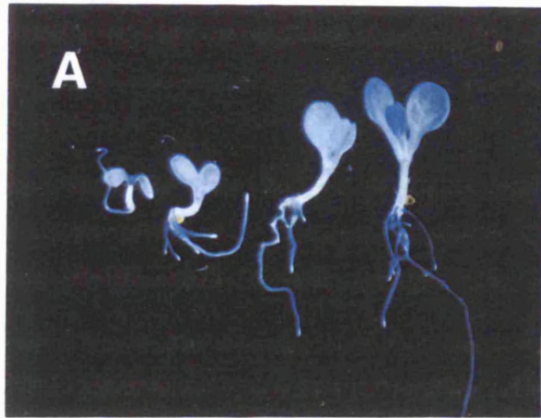
Figure 3.1 GUS activity distribution in AtEN101 plants.

The chart depicts the mean activity of the enzyme in different organs as detected fluorimetrically by conversion of MUG to MU (Section 2.10.1). Measurements were made in 4-8 samples for each organ (with 2-3 readings averaged for each sample). Sq-1: siliques containing embryos from early globular to early heart stages. Sq-2: siliques containing embryos from late heart-stage to torpedo stage. Emb3-1: early (green) cotyledonary stage embryos. Emb3-3: maturing cotyledonary stage embryos (white). The error bars show the standard deviations.

Figure 3.2 GUS activity in AtEN101 plants (II).

The reporter activity was detected by histochemistry as detailed in Section 2.9.2.

- A)** Seedlings 2, 3, 5 and 7 days after germination (grown in $1/2MS_{10}$).
- B)** 5 day-old seedling (grown in $1/2MS_{10}$).
- C)** Expression in root tip area of plant shown in B. Amplification 100 x.
- D)** Close-up of root in C. Amplification 200 x.
- E)** Three-week old plant grown in $1/2MS_{10}$.
- F)** Left: 3 and 2 days-old seedlings grown in $1/2MS$; Right: 3 days-old seedlings grown in $1/2MS_{10}$.
- G)** 3 day-old seedlings germinated and grown in the dark (Bottom) and under a normal day-night regimen (Top). With exception of the top-most region, the blue color seen throughout the hypocotyl is due to an optical effect during photography and not to GUS activity.
- H)** Mature leaf from a plant grown in soil.



3.2H). Similar findings were obtained with plants continuously grown hydroponically in 1/2MS₁₀ or transferred to 1/2MS up to 10 days prior to analysis.

Although no attempt was made to determine reporter gene activity in roots of mature soil-grown plants, roots of bolting or mature plants grown hydroponically in 1/2MS₁₀ or adapted to 1/2MS for up to 10 days exhibited GUS activity. The pattern of expression was similar to that found in seedlings, *i.e.* throughout most of them except in the meristems (Fig. 3.3A). However, there seemed to be variations in the intensity of staining along the length of a root and often no staining in older regions. Sectioning of the roots did not alter these staining patterns (data not shown) indicating that the absence or lower level of histochemical reaction in the older portions of the roots were not due to impermeability of the tissues.

Bolting was accompanied by an intense activity of the reporter in the growing inflorescence stem (Fig. 3.3B). This activity was considerably reduced in inflorescence stems of mature plants grown in soil or hydroponically in 1/2MS₁₀. In these mature plants, histochemical staining was not uniform along the inflorescences. Most of the detected GUS activity was concentrated in the apical portions of the stem and its attached flowers and young siliques (Figs. 3.3B, C, D and E), and also in axial buds (not shown). This activity was considerably higher in apical regions of plants grown hydroponically in 1/2MS₁₀ than on plants grown in soil (Fig. 3.3C and D). In general, histochemical staining decreased in intensity with age along the inflorescence, older portions not showing any or little GUS activity (*e.g.* Fig. 3.3B).

There seemed to be some correlation between the staining intensities of a given region of the stem and those of the flower and silique petioles attached to it (*e.g.* Figs. 3.3C, D and E). Petiole attachment regions often showed a higher degree of histochemically detectable GUS activity (*e.g.* Figs. 3.3E, G). The staining gradients observed in stems and petioles could reflect a genuine differential expression of GUS along their length. Alternatively, they could result from a variable degree of access of the histochemical reagent to different regions. To test for a possible age-related increase in impermeability of the stem and flower/silique petioles to the histochemical reagents, poorly staining stem regions were sectioned prior to histochemical detection. Much stronger staining was indeed observed in the region immediately adjacent to the cut surface (Fig. 3.3G) and GUS activity seemed to be present in all tissues (Fig. 3.3H). This indicates that the observed gradients of histochemical staining in stems and petioles are a consequence, at least partially, of permeability differences and do not necessarily reflect gradients of GUS activity.

In the reproductive organs the activity was found to follow a complex pattern. In younger bolting plants GUS activity was intense in flower buds including petals and sepals (Fig. 3.3B). The patterns varied with the growth conditions and age of the plant. In older plants no GUS activity was visible or was very limited in young flower buds, only progressively becoming visible in the filaments of the stamens, the tapetum and the pollen grains, prior to the

open flower stage (Figs. 3.3D, 3.4A). During flower opening and subsequent stages the enzyme remained active in pollen but not in the tapetum (Fig. 3.4B). In filaments the activity increased relative to earlier stages and then was reduced and eliminated (Fig. 3.3D). GUS was found in carpels from their early development in younger plants but not older plants (Figs. 3.3C, D, and E). In both situations the reporter enzyme was detectable in silique walls throughout the earlier stages of embryogeny up to the transition to torpedo stage (Figs. 3.3E and F).

Reporter activity was detected in all tissues of ovules and in embryos and embryo sacs from early globular stage persisting until late heart stage when it was no longer detectable (Fig. 3.4C-G). The enzyme was again active in embryos but not in the endosperm or other tissues of the developing seed from the walking-stick or early cotyledonary stages (Fig. 3.4H). In this second wave of expression GUS was initially detectable throughout the embryo but became progressively limited to the apical region of the axis and the cotyledons (Fig. 3.4H). Eventually it became restricted to the adaxial surface of the cotyledons before disappearing as the seeds completed their maturation.

In summary, the GUS enzyme could be found in most organs and tissues of EN101 plants. However, its expression was not constitutive being apparently excluded from the meristematic region of roots, and older regions of many organs. In other organs (*e.g.* stamens) the activity was found to be developmentally controlled, and/or influenced by physiological conditions. These results differed from those obtained by Topping *et al.* (1994), particularly with regard to mature plants. Also, during embryo development the pattern of expression was more complex than previously described.

3.2.4 Phenotypical characterization

Homozygous and hemizygous AtEN101 plants do not display any obvious phenotypic alteration relative to the wildtype. For example in mixed populations derived from self-pollination of hemizygous individuals it was never possible to classify the various plants with regard to their genotype as a function of their morphology or development. However, no detailed and quantitative studies were made of the plants and their responses to different growth conditions. It is then possible that the mutation might have phenotypic effects that went undetected. Alternatively the T-DNA insertion did not functionally disrupt a transcription unit, or such unit had a redundant role in the plant, at least under the growth conditions used.

3.3 Identification and characterization of the *bashful* mutant

3.3.1 Screening for mutations affecting seedling pattern and shape

The developmental continuity between embryo and seedling stages often signifies that mutations causing morphological aberrations in seedlings also affect embryogenesis (see

Figure 3.3 GUS activity in AtEN101 plants (III).

The reporter activity was detected by histochemistry as detailed in Section 2.9.2.

A) Roots of a mature, flowering plant grown hydroponically in 1/2MS₁₀.

B) Bolting plant grown in 1/2MS₁₀. Intense staining of the flower buds makes them difficult to distinguish from the background.

C) Terminal inflorescence of plant grown in 1/2MS₁₀.

D) Terminal inflorescence of plant grown on soil.

E) Inflorescence and young siliques (globular stage embryos) of a soil-grown plant.

F) Siliques from a same inflorescence stem. From globular (Top) to heart/torpedo stage transition (Bottom).

G) Inflorescence stem that was cut prior to staining.

H) Close-up of one of the termini visible in G.

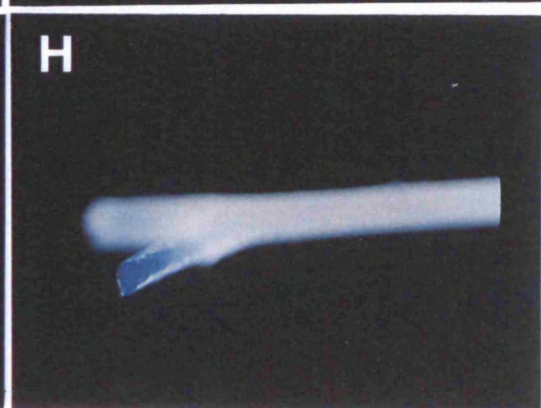
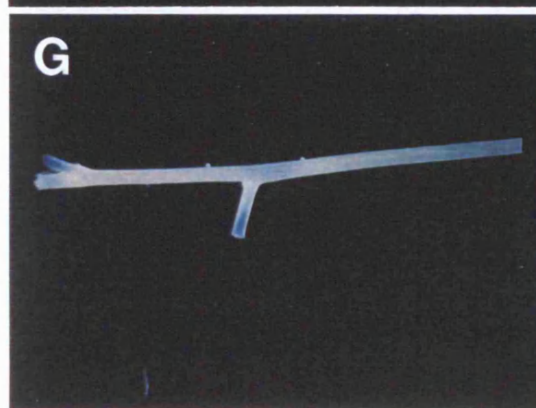
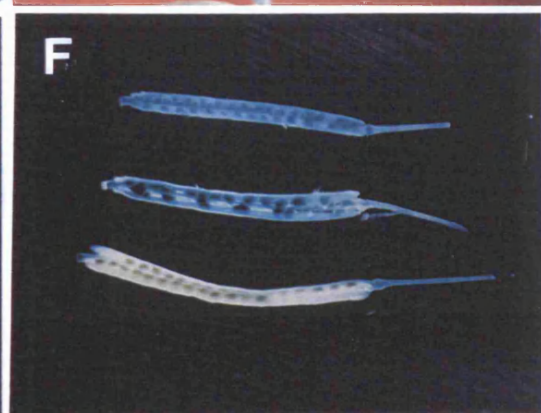
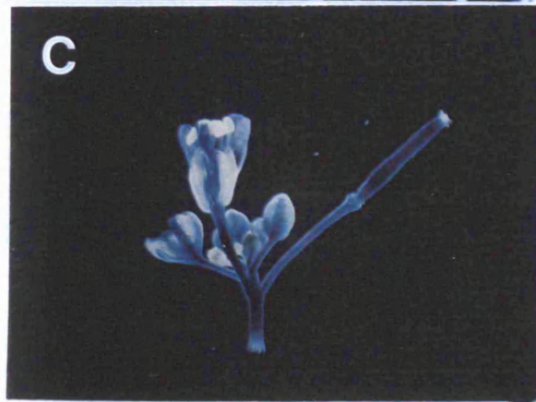
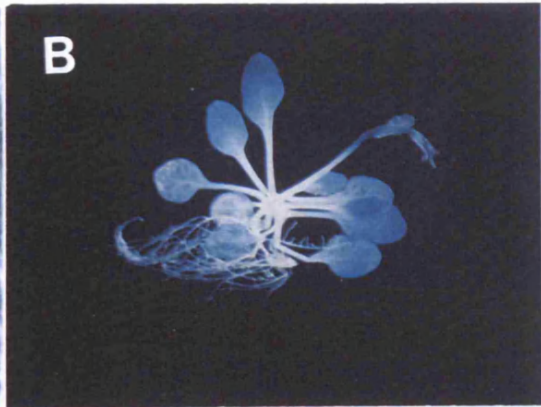
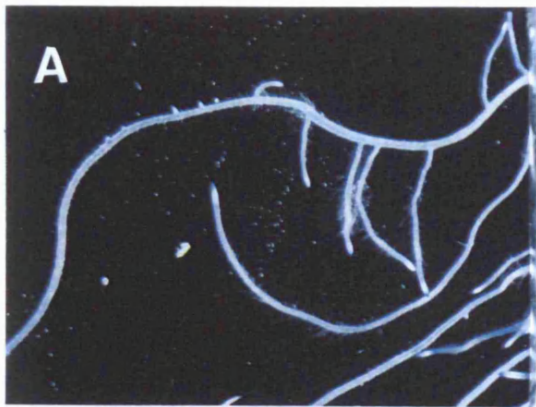
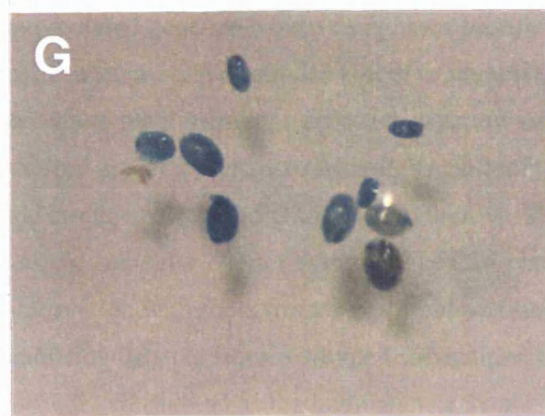
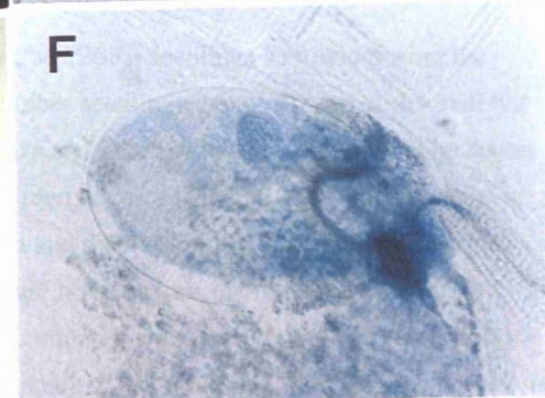
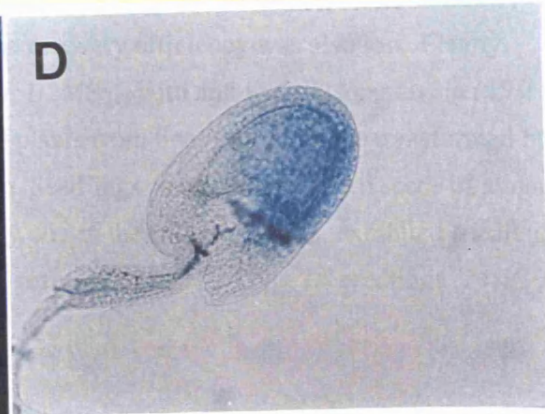
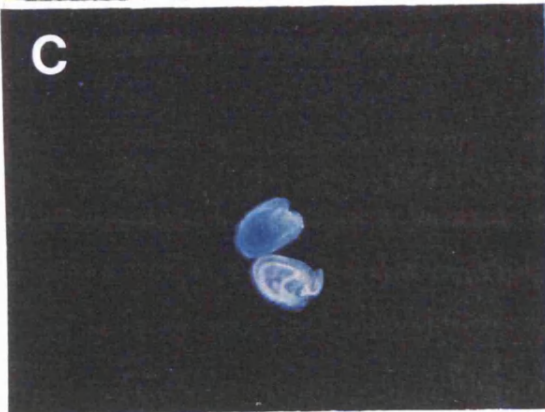
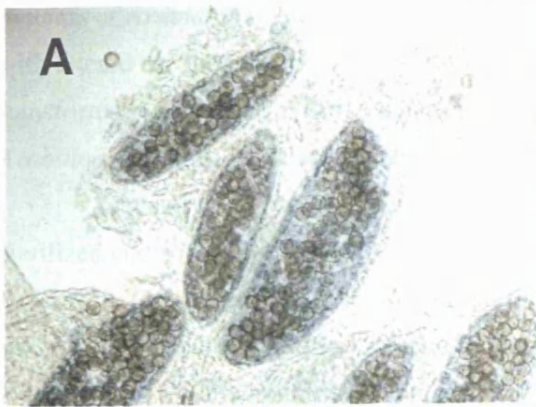


Figure 3.4 GUS activity in reproductive structures and embryos.

The reporter activity was detected by histochemistry as detailed in Section 2.9.2.

- A)** GUS Activity in pollen grains and the tapetum in flower buds. Amplification 400x.
- B)** Activity in pollen, stamens and stigma at open flower stage. Amplification 200x.
- C)** Early globular stage ovules from a same silique showing the time of activation of *gus* expression.
- D)** Globular stage ovule showing regional activation of *gus* including the embryo and suspensor.
- E)** Mid-globular stage ovule.
- F)** Ovule at globular/heart-stage transition. Most of the integuments and chalazal region have been removed to clearly see the embryo and embryo-sac.
- G)** Collection of ovules at different stages of development up to late torpedo stage.
- H)** Cotyledonary/maturation stage embryos showing variation in the patterns of GUS activity.



Section 1.2.3). This observation was first exploited by Jürgens and co-workers to isolate pattern mutants of *Arabidopsis* (Jürgens *et al.*, 1991; Mayer *et al.*, 1991). To identify and isolate mutants with altered seedling morphology that might also be embryo-defective resulting from transformation with p Δ gusBin19 a partial screening of the available library of transgenic *Arabidopsis* lines (Clarke *et al.*, 1992) was performed.

Seeds of a total of 111 primary transformants were cold-treated for 10-14 days, surface-sterilized and sown. Initially batches of seeds were plated on 1/2MS₁₀ agar supplemented with kanamycin (50 µg/ml). However, for a high proportion of the lines it was not possible to recover seed-setting plants because seeds failed to germinate or seedling-morbidity likely caused by the kanamycin. To overcome this difficulty, other batches of T2 seeds were grown on perlite supplemented with a nutrient salt medium but the recovery efficiency was also low. Finally, seeds from primary transformants were plated on 1/2MS₁₀ with and without kanamycin (35 µg/ml). It was then possible to obtain seed-setting plants from lines that might be transformed but over-sensitive to kanamycin in the T2 generation. Seedlings were obtained from seeds of about 60% of the primary transformants screened. Only one of these lines, P26K4, exhibited seedlings with defective shape. Of 57 seeds of P26K4 that germinated in 1/2MS₁₀, 14 seedlings (24.6%) showed the defect described below.

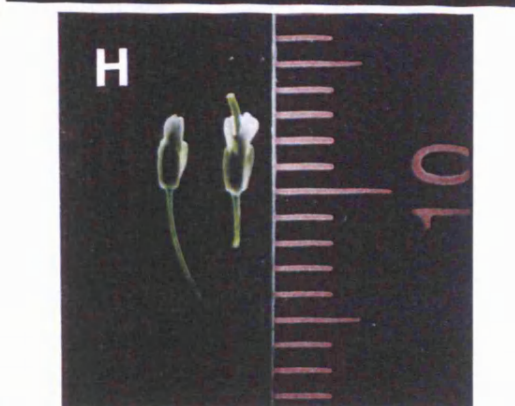
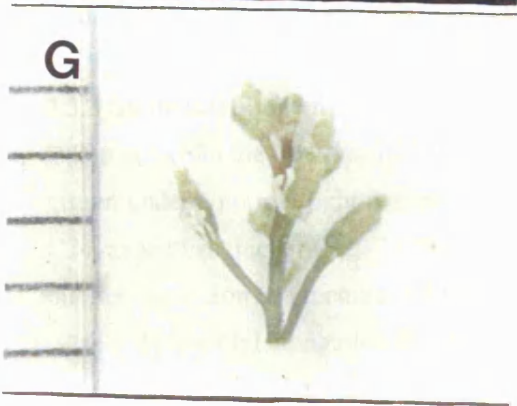
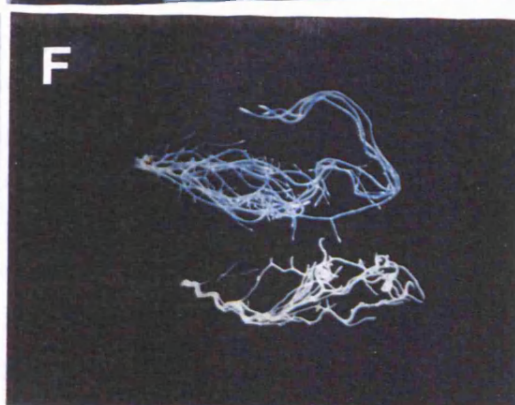
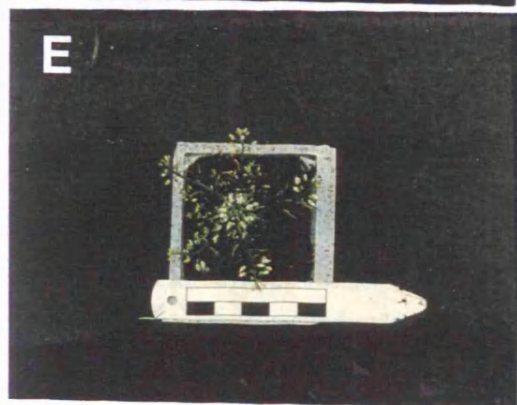
3.3.2 The *bashful* phenotype

The distinctive characteristics shared by some of the P26K4 seedlings as found during the screening included slower growth relative to the other seedlings, being stunted with small but sturdy, thicker round leaves almost identical in size and in shape to the cotyledons. The leaves formed a tight rosette and were curled and dark-green in color (Fig. 3.5A, B). This phenotype has been shown to be monogenic, stably inherited and displayed only by plants homozygous for the mutation (see below, Section 3.3.3). The mature homozygous plants grown in soil are strikingly short, rarely exceeding 6 cm in height compared with the >30 cm for their wildtype counterparts (Fig. 3.5C, D), and because of this apparent "shyness" and dwarfing the phenotype and mutated gene have been designated *bashful*. The mutant plants usually display a central inflorescence stem of similar length to several other (5-6) radially oriented stems (Fig. 3.5E), the whole plant forming a compact structure with a bushy appearance. This is indicative of reduced apical dominance. Although smaller in length, stems, leaves and cotyledons are thicker and stronger compared with those in their wildtype background, C24. The roots of homozygous *bashful* were also shorter than those of C24 plants of the same age. Figure 3.5F illustrates this observation with roots from 5 week-old seedlings of *bashful* (non-stained) and of EN101 (indistinguishable from wildtype C24; stained for GUS activity). All these attributes are formally considered as the hallmarks of dwarf mutants as defined by Feldmann (1991).

The reproductive organs of homozygous *bashful* plants are considerably smaller in size relative to the wildtype. A developing silique at the open flower stage is just about 1 mm long

Figure 3.5 The *bashful* phenotype (I).

- A) *bashful* homozygous seedling growing on 1/2MS₁₀.
- B) *bashful* homozygous seedling growing on soil.
- C) Same age wildtype C24 plants (Left) and an homozygous *bashful* plant (Right).
- D) Well-grown, mature *bashful* plant. The bar scale is in cm.
- E) Top view of a mature *bashful* plant. The bar scale is in cm.
- F) Root systems from 5 week-old EN101 plant (Top; following histochemical detection of GUS) and *bashful* plant (Bottom).
- G) *bashful* inflorescence (bar scale in mm).
- H) Wildtype C24 mature flowers (bar scale in mm).



(3-4 mm in C24) and the stamens are proportionally small (Fig. 3.5G, H). Silique elongation during development is reduced and at maturity they do not exceed 4 mm. In contrast, C24 mature siliques reach up to 15 mm in length. Although the inflorescence stems are very short, each accumulates numerous flowers and siliques along their length separated by very short internodes (*e.g.* Fig. 3.5D). The life cycle of these plants including the flowering period is longer than for wildtype C24 lasting on average over 90 days after germination. In spite of the relatively numerous seeds produced, these plants show low fertility, *i.e.* only a fraction of the seeds are viable.

Seeds and mature embryos of homozygotic *bashful* plants were inspected to determine if the mutation affected their development. Often mature seeds had a distorted appearance as illustrated in Figure 3.6A. The altered seed forms corresponded to morphologically aberrant embryos within, as can be seen in Figure 3.6B. These alterations could be recognized from early cotyledonary stage (Fig. 3.6C). The morphologically aberrant embryos displayed a variety of forms and sizes (*e.g.* Fig. 3.6B) but upon germination the resulting plants had the common *bashful* trait. This suggested that the observed abnormalities were not due to a genetic factor. Also, importantly, the seeds in heterozygous plants have a quite normal appearance. The small size of the siliques could be a factor in determining the shape of the seeds and embryos particularly as tens of ovules can be found within each. It therefore seems that shape-modified seeds result from physical constraints imposed on them by the small size and the cramped conditions of the siliques, as can be seen in Figure 3.6D. This observation may also explain the low fertility of the homozygous plants. It was noted that viable seeds tended to be the largest. It is possible that those developing ovules that fail to grow sufficiently due to space constraints either fail to complete all required stages of development and/or to accumulate enough nutrient reserves to be able to germinate under normal conditions.

Heterozygous *bashful* plants are morphologically similar to wildtype C24 plants but empirical observations indicate that they tend to be smaller and accumulate anthocyanins in their rosette leaves as exemplified by the plant in Figure 3.6E. However, no systematic and quantitative analysis of these plants has been carried out.

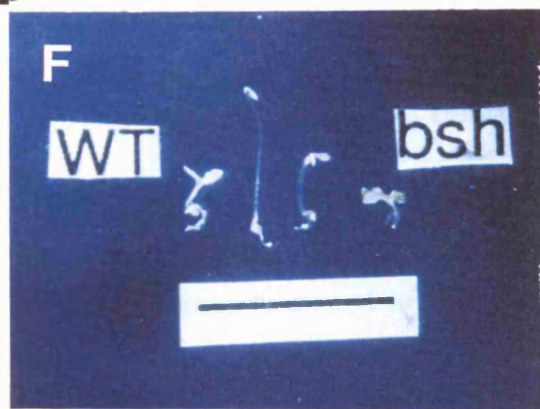
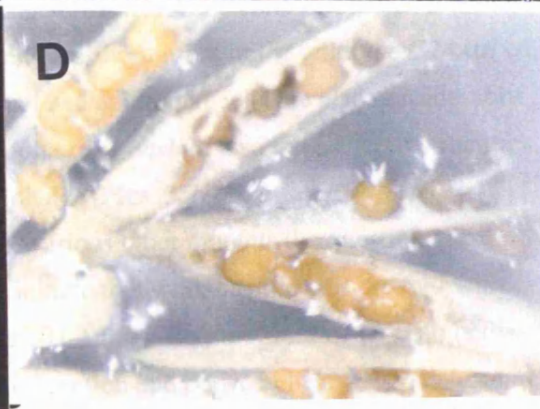
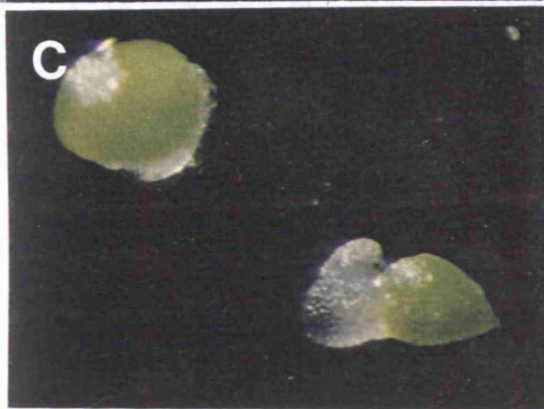
3.3.3 Skotomorphogenic development in *bashful*

When grown in the dark *bashful* plants display elongation of their hypocotyls relative to those grown under a normal light regime. However these hypocotyls are shorter than in dark-grown C24, as well as thicker (Fig. 3.6F). A properly formed hook is not present and there is opening but not expansion or greening of the cotyledons (Fig. 3.6F).

Hypocotyl elongation is not observed when *bashful* seedlings are grown under red, far-red or blue light, unlike C24 seedlings (Fig. 3.6G). One week-old seedlings grown under far-red light showed open, light-green cotyledons and the first pair of leaves. Open cotyledons and the first pair of leaves were also present in seedlings grown exposed to red light but they were

Figure 3.6 The *bashful* phenotype (II).

- A) Seeds from a same silique of an homozygous *bashful* plant. A seed from a wildtype C24 plant is shown at the bottom left corner.
- B) Embryos dissected from the seeds shown in A. The embryo from the wildtype seed is at the bottom right corner.
- C) Seeds at early cotyledonary stage.
- D) Mature seeds in diehescant siliques.
- E) Heterozygous *bashful* plant.
- F) Developmental dependency on light conditions. From left to right: 3 day-old C24 plants grown with white light, and in the dark. *Bashful* plants grown in the dark and with white light. The bar is 1 cm.
- G) Developmental dependency on light conditions. *bashful* and C24 plants were germinated and grown for 7 days under different light conditions. The 3 *bashful* and the two wildtype green plants on the right were grown under white light. On the left are plants grown under other light conditions: Top row- 3 *bashful* and C24 wildtype seedlings grown with red light; Center row: *idem* grown with far-red light; Bottom row: 3 C24 and 3 *bashful* seedlings grown with blue light.



etiolated. A blue light regime resulted in cotyledons and a primary pair of leaves that were expanded and larger than in those plants grown under red and infrared light. These organs were also green, albeit not as green as in seedlings germinated under normal white light conditions.

3.3.4 *bashful* is not allelic to *diminuto*

A number of mutants with *bashful*-like phenotypes have been described including *dwarf1 (dwf1)/diminuto (dim)* (Takahashi *et al.*, 1995; Kauschmann *et al.*, 1996), *cpd* (Szekeres *et al.*, 1996), *det2* (Chory *et al.*, 1991) and, more recently, *dwarf4 (dwf4)* (Azpiroz *et al.*, 1998). These mutants are characterized by defective cell elongation. Microscopic examination of the hypocotyl and roots of *bashful* seedlings showed that cell expansion is also affected in this mutant (Fig. 3.7A-C). Although no detailed characterization of the cells was made, it is clear that in both instances, cortical and epidermal cells are shorter than in the wildtype C24 (to which *AtEN101* is similar). In the hypocotyl the cortical and epidermal cells are also considerably wider than in C24 (Figs. 3.7C), whereas the root hairs of *bashful* also have altered structures, being shorter than in C24 and terminating in globules (Figs. 3.7A, B).

Given the phenotypical similarity to the various other mutants the question arose of the possible allelism of *bashful* to any of them. To initiate this work *bashful* was crossed with *diminuto* plants (kindly provided by Professor N.-H. Chua). Crosses were made both between homozygous plants of the two mutants in either direction and between heterozygous *bashful* plants (female) and homozygous *diminuto* plants (male). F1 seeds from all crosses were germinated and in all cases the plants did not have a dwarf phenotype. Quite the contrary, they were exceptionally strong with profuse rosettes demonstrating their hemizygosity. This proved that the two mutations are not allelic.

To establish if the *bashful* and *diminuto* loci were linked a number of F2 seeds derived from F1 of crosses of homozygous were germinated and the seedlings classified according to the presence or absence of a dwarf phenotype (Table 3.1). The results showed the absence of a tight linkage between the two loci. However, there is a significant deviation from the expected ratio of 43.8% of dwarves. This could be due to a reduced viability of seeds enclosing embryos of dwarf plants or to linkage of the two loci. The results in Table 3.3 indicate that dwarfism does not affect viability.

	Dwarf	Non-dwarf
F2 Plant 1	11 (14.5%)	65 (85.5%)
F2 Plant 2	172 (34.1%)	331 (65.9%)
F2 Plant 3	138 (39.4 %)	212 (60.6 %)
Totals	321 (34.6 %)	608 (65.4 %)

Table 3.1 Phenotypic distribution of F2 plants from homozygous crosses of *bashful* and *diminuto*.

3.3.5 The *bashful* phenotype can be rescued by brassinosteroids

Several dwarf mutants like *dim*, *dwf4*, *cpd*, *det*, *cop*, *fus*, and *axr2* can be rescued by feeding with brassinosteroids (Kauschmann *et al.*, 1996; Szekeres *et al.*, 1996; Azpiroz *et al.*, 1998). Homozygous *bashful* plants were germinated and grown in solid 1/2MS₁₀ supplemented with 0.1 mM or 1 mM of either epibrassinolide, a brassinosteroid, or gibberellic acid (GA₃). The gibberellic acid had no effect on the plants. In contrast, seedlings grown in 1 mM epibrassinolide, and to lesser extent those in 0.1 mM, responded to the presence of the hormone with alterations in leaf shape and extension of petioles (Fig. 3.7D-F).

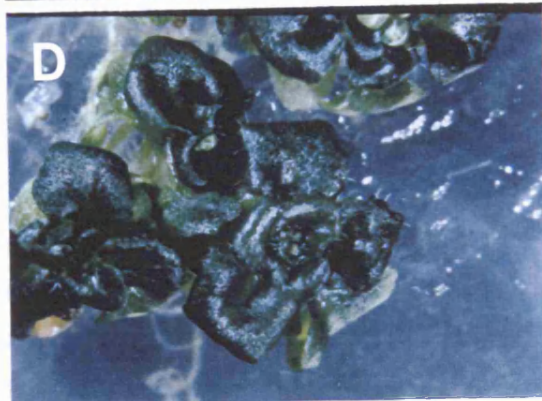
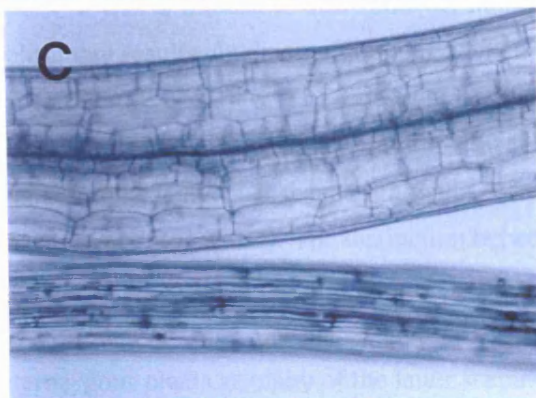
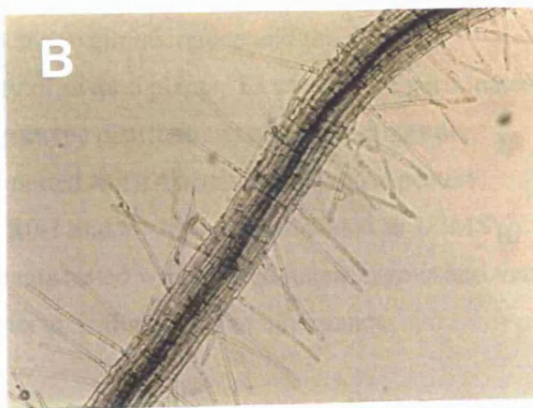
3.3.6 Linkage of the *bashful* phenotype to a T-DNA locus

Twenty-eight of the P26K4 T2 seedlings were transferred to soil including 7 displaying the phenotype. Seeds were separately collected from each potted plant. These mutant plants showed the typical dwarf phenotype, described above (Section 3.3.2), throughout their growth. Several of the non-dwarf plants although not showing any obvious morphological defect were clearly smaller and particularly rich in anthocyanin pigmentation in the rosette leaves compared with their other non-mutant siblings and wildtype plants (Fig. 3.6E). These plants were suspected of being heterozygous for the mutation detected. To test this possibility and to obtain information on the segregation ratio of the mutant phenotype, T3 seeds from two of these plants, plants #5 and #6, were plated on 1/2MS₁₀ supplemented with kanamycin to 50 µg/ml (Table 3.2). None of the kanamycin-sensitive plants showed the dwarf phenotype. Seeds from 3 of the dwarf plants were also plated on kanamycin-containing 1/2MS₁₀. All seedlings displayed the phenotype and were kanamycin-resistant although some seemed to be affected by the presence of the antibiotic as they would partially bleach (Fig. 3.7G). These results indicated that the dwarf phenotype was heritable, stable (being similar in different seedlings) and co-segregated with kanamycin resistance. However, it was not possible to accurately determine the segregation ratios in the progeny of heterozygous plants also because it was not known if sensitivity to kanamycin could have prevented germination of some seeds.

Family	Kan ^r Dwarf	Kan ^r Non-dwarf	Kan ^s (all non-dwarf)	Non- germinated
Plant 5	4 (30%)	4 (30%)	5 (38%)	9
Plant 6	18 (32%)	34 (61%)	4 (7%)	unknown

Table 3.2 Phenotype segregation among progeny of T3 heterozygous *bashful* plants.

The numbers and percentages of sibling plants of two families categorized according to resistance to kanamycin and the dwarf phenotype are shown.



Having established a likely link between kanamycin resistance and the dwarf phenotype, to confirm the suspected heterozygosity of several plants, T3 seeds were germinated in 1/2MS₁₀ lacking kanamycin. The observed phenotype distributions of the seedlings are shown in Table 3.3. Several of the seedlings were potted and T4 seeds from two suspected heterozygous plants derived from plant 10 (plants 10-1 and 10-2) were also plated in 1/2MS₁₀ (Table 3.3). It was clear from the results that all plants tested were indeed heterozygous and that the dwarf phenotype was associated with a single locus with mendelian inheritance (approximately 3:1 ratio).

It was of interest to determine the number of T-DNA loci in the mutant but the preliminary results obtained with plants 5 and 6 (see above Table 3.2) were not clear in this respect. To this end T4 or T5 seeds from several families were grown on 1/2MS₁₀ (35 µg/ml kanamycin) and the seedlings scored for resistance to the antibiotic. As expected, all dwarf plants were kanamycin-resistant although several were affected in their growth rate and partial bleaching (*e.g.* Fig. 3.7G). The distinction between wildtype and heterozygous seedlings was not always obvious and only the more reliable results obtained with seeds of plants A, E, 14 and N-1 are listed in Table 3.4. Because of the difficulties in distinguishing wildtype from heterozygous plants as many of the latter seemed to be affected by kanamycin, the experiment was repeated after lowering the concentration of the antibiotic in the media to 25 µg/ml. This strategy was fruitful and the phenotype distributions are presented in Table 3.4.

The results were not consistent as evidenced by substantial variations in the proportions of the various types of seedlings even when relatively large numbers of seedlings were involved. However, all dwarfs were kanamycin-resistant confirming their linkage to a T-DNA tag. The observed proportional variations were very unlikely to be due to systematic classification errors and rather to the presence of additional segregating resistance markers. Corroborating this interpretation was the identification of a new mutant phenotype in the P26K4 line represented in two plants. These plants, including plant NM-1 (Table 3.4) were generally characterized for being very robust (data not shown). Whether this mutation was linked to an additional T-DNA locus was unknown but possible.

To reduce or eliminate the likely presence of T-DNA loci not linked to the dwarf mutation, heterozygous seedlings from 3 of the families showing a closer to expected distribution of kanamycin-resistant plants for a single T-DNA were selected (plants E2-2, E2-3 and P3-9; Table 3.4) and transferred to soil. Only the progeny of these plants was utilized for subsequent work.

3.3.7 Presence of *gus* reporter genes in *bashful*

To detect the presence of fusion transcripts (or proteins) involving *gus* and the affected gene in *bashful* a histochemical approach was used. No GUS activity was detected in any organ at different developmental stages of the homozygous and heterozygous plants tested, even when

Family	Non-dwarf	Dwarf	Non-germinated
Plant 1	10	5	0
Plant 2	8	4	0
Plant 3	11	7	0
Plant 4	17	6	0
Plant 7	17	4	0
Plant 8	15	9	0
Plant 9	21	12	0
Plant 10	25	5	2
Plant 11	47	18	unknown
Plant 12	22	10	0
Plant 10-1	22	11	unknown
Plant 10-2	32	8	unknown
Totals	247 71.4%	99 28.6 %	unknown

Table 3.3 Dwarfism in the progeny of suspected heterozygous plants.

T3 and T4 seeds of 12 non-dwarf plants but displaying high anthocyanin content in leaves and apparent smaller size than wildtype plants were grown in 1/2MS₁₀. The numbers of dwarf seedlings and those not displaying the phenotype (wildtype/heterozygous) were counted.

Family	Wildtype	Heterozygous	Dwarf
Plant A	5 (29.4%)	9 (52.9%)	3 (17.6%)
Plant E	4 (33.3%)	6 (50.0%)	2 (16.7%)
Plant 14	5 (26.3%)	9 (47.4%)	5 (26.3%)
Plant NM-1	0	38 (95.0%)	2 (5.0%)
Plant P3-1	0	30 (81.0%)	3 (9.0%)
Plant P3-9	13 (30.2%)	19 (44.2%)	11 (25.6%)
Plant P3-12	2 (4.2%)	32 (66.7%)	14 (29.2%)
Plant E2-2	20 (30.7%)	31 (47.7%)	14 (21.5%)
Plant E2-3	9 (24.3%)	17 (45.9%)	11 (29.7%)
Plant E2-7	29 (35.4%)	37 (45.1%)	16 (19.5 %)
Plant 11	9 (9.9%)	61 (67.0%)	21 (23.1%)

Table 3.4 Segregation of kanamycin-resistance and dwarf phenotypes.

T4 or T5 seeds of the P26K4 transgenic line were germinated in 1/2MS₁₀ supplemented with 35 µg/ml (plants A, E, 14, NM-1) or 25 µg/ml kanamycin and the seedlings scored for dwarfism.

the material was incubated in reaction solution for 48 hours. The lack of detection of GUS in these plants could have several explanations including the absence or a very low level of expression, or the production of an enzymatically-defective protein, or even to the reverse orientation of *gus* relative to a transcription unit where it might have integrated.

To test the presence of the *gus* reporter in the *bashful* genome a Southern hybridization experiment was conducted. Genomic DNA was prepared from a pool of homozygous plants and aliquots were digested with one of several restriction enzymes known to have sites within the T-DNA region of p Δ *gus*Bin19. The enzymes chosen were *Bam*H I, *Eco*R I, *Eco*R V, *Hinc* II, *Hind* III, *Ssp* I and *Sst* I. The digested DNA aliquots were resolved in an agarose gel, transferred to a nylon membrane by capillarity and fixed, as described in Section 2.5. The dry blot was hybridized with a probe prepared by random oligonucleotide priming (Section 2.8.2) from the *gus* gene and labelled with [α -³²P]dCTP. For preparation of the probe the *Bam*H I/*Eco*R I insert of pGUS containing the entire *gus* gene (Section 2.1.2.5) was used. Autoradiography of the hybridized Southern blot revealed a number of fragments containing *gus* sequences in the genome of *bashful* plants (Fig. 3.8). These fragments are listed in Table 3.5.

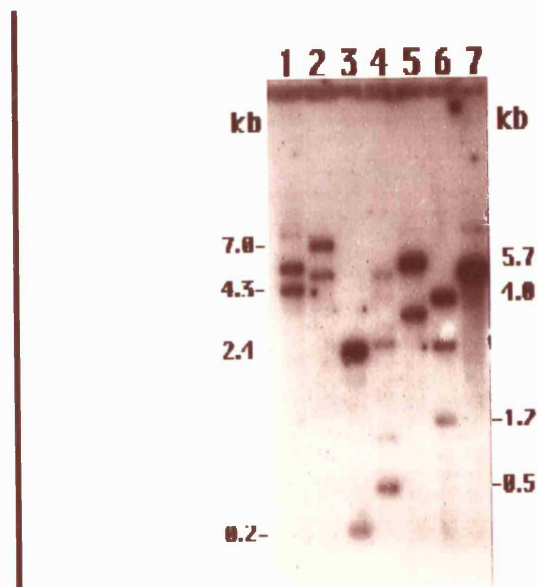
The Southern blot results indicated that there were two *gus* genes in *bashful*. This was suggested by the two bands observed in *Bam*H I, *Eco*R I and *Hind* III digests, taking in consideration the restriction enzyme map of the T-DNA in p Δ *gus*Bin19. This is particularly true of *Eco*R I and *Hind* III for which digestion should not be prevented by methylation at their unique sites in the T-DNA. Also, the intensity of the 2.4 kb and 0.2 kb *Eco*R V bands indicated that they contained twice the fragments expected from a single *gus* gene. In fact, in another Southern analysis experiment it was possible to distinguish two bands of about 2.4 kb (data not shown) confirming the double nature of the band observed. This also indicated that at least one of the hybridizing fragments had an altered structure. The strong hybridization signal detected in the *Sst* I digest likely resulted from a double band. More difficult to

<i>Bam</i> H I	<i>Eco</i> R I	<i>Eco</i> R V	<i>Hinc</i> II	<i>Hind</i> III	<i>Ssp</i> I	<i>Sst</i> I
5.6 kb	7.0 kb	2.4 kb	5.1 kb	5.7 kb	4.0 kb	5.0 kb
4.3 kb	5.0 kb	2.4 kb	4.1 kb*	3.2 kb	2.4 kb	-
-	-	1.7 kb*	2.5 kb	-	1.2 kb	-
-	-	1.3 kb*	1.4 kb*	-	-	-
-	-	0.2 kb	1.1 kb*	-	-	-
-	-	-	0.5 kb	-	-	-

Table 3.5 Restriction enzyme fragments of the *gus* gene in the *bashful* genome.

The sizes of the genomic fragments hybridizing to a probe derived from a complete *gus* gene in the Southern blot shown in Figure 3.8 are listed. The asterisks mark faint hybridizing bands. Some fainter bands are not listed.

A)



B)

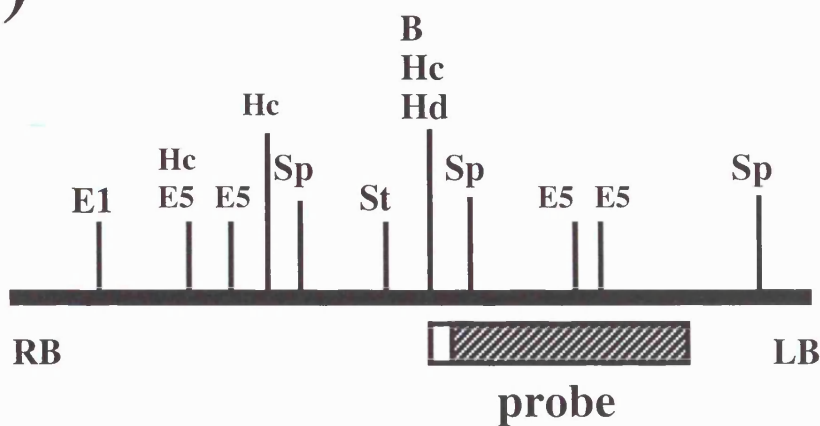


Figure 3.8 Southern blot detection of *gus* in *bashful* plants.

A) A blot prepared from DNA digested with various restriction enzymes was hybridized with a probe prepared from *gus*. Shown is an autoradiograph of the hybridized blot. Lane 1: *Bam*H I; Lane 2: *Eco*R I; Lane 3: *Eco*R V; Lane 4: *Hinc* II; Lane 5: *Hind* III; Lane 6: *Ssp* I; Lane 7: *Sst* I.

B) Restriction map of the T-DNA in pΔ*gus*Bin19. The position of the *gus* gene from where the hybridization probe was prepared is shown by the hatched box.

interpret were the restriction patterns for *Hinc* II and *Ssp* I, both of which enzymes are not affected in their activity by site methylation. No *Hinc* II sites exist in the *gus* gene. Thus, the complex pattern and the small sizes of *Hinc* II fragments obtained show that the T-DNA was not intact in the left border region(s). The 1.2 kb *Ssp* I fragments correspond in size to the *Ssp* I fragment at the 3' end of *gus* in the T-DNA of p Δ *gus*Bin19. Because the intensity of hybridization of *Sst* I fragments did not indicate the presence of double bands, it is likely that at least one of the two *gus* genes is not intact in *bashful*. The *Eco*R V pattern suggested that the central and 3' region of the genes should be intact. Therefore, it is likely that structural alterations/truncations affecting the the T-DNA should be located in the 5' region and downstream from *gus*. This is also borne by the observation that the 1.9 kb *Sst* I segment corresponding to most of *gus* and the left border region of p Δ *gus*Bin19 is absent in *bashful*.

In all digests, a number of faint bands were visible, some of which are listed in Table 3.5. It is possible that these derived from truncated *gus* segments present in the genome. The emerging picture of the T-DNA in *bashful* indicates a complex organization where no complete *gus* genes may be present. The absence of a functional reporter gene could explain the lack of detection of GUS activity in the mutant.

3.4 Summary

Fluorimetric and histochemical studies of GUS activity in homozygous AtEN101 plants showed that the enzyme was active in most tissues and organs (Section 3.2). Expression of *gus* was, however, developmentally regulated and was notably absent from some tissues. In particular, histochemical analysis failed to reveal the presence of the enzyme in root meristematic regions. However, strong activity was observed in shoot apical and axial meristems. During embryogenesis two waves of GUS activity were observed. During the first, the enzyme was detected throughout the ovule whereas in the second it was limited to the cotyledonary stage embryos. Spatial control of expression was also observed in the latter embryos with the enzymatic activity initially found throughout the axis and cotyledons being progressively limited to the adaxial region of the latter.

Considerable differences, particularly involving mature plants, were observed relative to a prior characterization of *gus* expression in the primary transformant and its seeds (Topping *et al.*, 1994). However, the patterns of expression observed were consistent across several generations. It is suggested that the observed disparities resulted from physiological peculiarities of the primary transformant due to the transformation procedure and/or culture conditions that affected *gus* expression. Indeed, during the experiments described it was observed that reporter gene expression was influenced by the growth conditions of the plants.

No obvious specific phenotypic trait associated with the mutation was detected in AtEN101 plants, regardless of their being hemizygous or homozygous with regard to the T-DNA. This indicated that if a gene had been mutagenized by insertion of the promoter trap, the production and activity of its encoded product was not affected by the T-DNA. Alternatively, the putative mutated gene was functionally redundant, at least under the growth conditions used.

Also reported in this Chapter was the screening of T2 seedlings from a number of primary transformants of p Δ *gus*Bin19 for shape and pattern mutants (Section 3.3.1). The preliminary characterization of a shape mutant isolated in this manner is reported (Section 3.3.2). Homozygous mutant plants displayed a typical dwarf phenotype and the trait was designated *bashful*. The mutation affected cellular dimensions and skotomorphogenic development. The mutation seems to be semi-dominant as heterozygous plants show some alterations relative to the parent wildtypes.

Co-segregation analysis showed that the mutation in *bashful* was linked to a kanamycin-resistance marker present in the T-DNA (Section 3.3.6). However, no reporter gene activity was observed. Southern blot analysis demonstrated the presence of *gus* gene sequences in the genomes of homozygous *bashful* plants. The results of the analysis suggested that two reporter genes are likely present but with structural alterations. A consequence of such alterations might be the absence of a functional gene thus explaining the lack of reporter activity in the mutant.

Crossing of *bashful* with *dim/dwfl*, a previously described dwarf mutant with similar phenotype (Kauschmann *et al.*, 1996; Takahashi *et al.*, 1995) showed that they do not carry defective alleles of the same gene. Preliminary rescue experiments were performed with gibberelin and with brassinosteroids. A positive response to brassinosteroids suggests that the mutant could be affected in the production of these growth regulators.

Chapter 4

Mapping transcripts to the T-DNA insertion region

4.1 Introduction

Determining if a gene is located in a tagged region of the genome requires isolation and characterization of the latter. The expression of *gus* in AtEN101 implied the activity of a promoter, possibly one of a genuine native plant gene, close to the T-DNA insertion site. This chapter describes the characterization of the tagged region of the transgenic genome and the identification of transcripts presumably mapping to it.

Earlier studies of AtEN101 by other workers had established that a single T-DNA copy was present per haploid genome of this line (Topping *et al.*, 1994). Topping *et al.* had also demonstrated the presence in the AtEN101 transgenic plants of a T-DNA fusion transcript of 3.5-4.0 kb by Northern hybridization analysis (Topping *et al.*, 1994). The junction region of the T-DNA left border and the adjacent native genome had been isolated following amplification by IPCR from a *Bam*H I fragment containing it, and was subsequently cloned (Dr. Topping, personal communication). Assuming that the T-DNA LB region in AtEN101 was sufficiently intact for the fusion transcription unit to have its 3' end at the *nos* terminator downstream from the *gusA* gene, the 5' end of the corresponding transcript was anticipated to be located about 0.7-1.2 kb upstream from the T-DNA left border. It should therefore be located within the native plant genome. Thus, the availability of native genomic sequences present in the cloned IPCR product could permit the investigation of the expected overlap of a native transcription unit(s) with the T-DNA insertion point by means of Northern blotting hybridization and the isolation of corresponding cDNA clones.

As described below, the cloned IPCR product containing the T-DNA/native DNA junction region was structurally characterized and used for detection of native transcripts by Northern blot hybridization (Section 4.2). However, the sequence characteristics of the IPCR product and its failure to hybridize to transcripts of wildtype *Arabidopsis* indicated its inadequacy for the isolation of cDNAs. For this reason, a genomic library of wildtype *Arabidopsis* in λ bacteriophage was screened by plaque hybridization to obtain clones of the region tagged in AtEN101 (Section 4.3). A total of four clones with similarity to the IPCR product were recovered. Longer genomic segments from the tagged region, derived from one of the genomic clones, were then utilized to screen a wildtype cDNA library. This permitted the identification and isolation of a total of eight cDNA clones with sequence similarity to the tagged region (Section 4.4).

4.2 Characterization of the T-DNA tagged region of AtEN101

4.2.1 Selection of an IPCR clone

The IPCR fragment containing the LB junction of the T-DNA had been cloned into the vector pCRTMII (Invitrogen, Dr. Topping, personal communication). The pCRTMII plasmid is a 3.9 kb-long T/A-type vector designed for easy cloning of *Taq* polymerase-amplified PCR products. Its cloning site is flanked by polylinker sequences including the only *Bam*H I and *Eco*R V sites and, on each side, the two *Eco*R I sites. The derivation of the IPCR clones is schematized in Figure 4.1. Two of the clones known to contain a single insert copy of the IPCR product, pIPCR1 and pIPCR2, were kindly offered by Dr. Topping. The orientation of the inserts in these plasmids was unknown. To establish their orientations, DNA of the two plasmids was prepared as described in Section 2.4.1. Samples of these were digested with *Bam*H I and then analyzed by agarose gel electrophoresis (Fig. 4.2A). The plasmid pIPCR1 was digested into two fragments of 5.9 kb and 0.4 kb (the latter is not visible in Fig. 4.2A). In pIPCR2 the digestion released an insert fragment of about 1.9 kb, presumably containing the T-DNA junction, from the vector and the *nos*-terminator segment. Schematic interpretations of the structures of the two plasmids are shown in Figures 4.2B and 4.2C. Because of the more convenient orientation of its insert, pIPCR2 was chosen for further analysis.

4.2.2 Sequence at the T-DNA border

Prior to the use of pIPCR2 to probe the existence of wildtype transcripts overlapping the T-DNA/native genome transition region it was of interest to determine the position of the junction within the plasmid insert. Barring a complex T-DNA arrangement, which was not anticipated given that a single T-DNA copy was thought to be present in AtEN101 (Topping *et al.*, 1994), the insert in pIPCR2 should have the structure shown in Figure 4.2C. Therefore, to obtain the sequence at the presumptive junction, sequencing was performed with two specific oligonucleotide primers for the T-DNA border, LB-63 and LB-267, by the dideoxy method as described in Section 2.5.4.1. As their names indicate, the two primers map to positions 63 and 267 bp downstream from the (intact) T-DNA left border and are oriented towards it (Dr. Topping, personal communication). The sequence of this region was found to be nearly identical to the corresponding T-DNA sequence in the original plant transformation vector pΔ*gus*Bin19 (Fig. 4.3). Only the terminal 7 bp at the left border were missing in pIPCR2.

4.2.3 Building a restriction and features map for pIPCR2

To facilitate the study of the structure of the genomic region of AtEN101 represented in pIPCR2 a simple restriction map of it was constructed. To this end the plasmid was digested with *Bam*H I and the DNA was fractionated in an agarose gel. The relevant 1.8 kb *Bam*H I fragment of the insert was isolated. Following purification, this fragment was subjected to

Figure 4.1 Schematic representation of the cloning of the LB junction region of AtEN101.

(J.F. Topping, personal communication).

- 1) Genomic DNA was digested with the restriction enzyme *Bam*H I.
- 2) The linear fragments were ligated under conditions favouring circularization.
- 3) The segment containing the junction was subjected to amplification by IPCR with primers for the known T-DNA sequence.
- 4) The amplified linear fragments were ligated into the T/A-type vector pCRTMII (Invitrogen). The two possible orientations of the insert are shown. Small arrows represent oligodeoxynucleotide primers used for PCR amplification. The left border (LB) and the T-DNA-derived segments are indicated by the white and black boxes, respectively. Other T-DNA sequences are represented by a thick bar. Restriction enzyme sites are labelled: *B*: *Bam*H I; *E1*: *Eco*R I; *E5*: *Eco*R V.

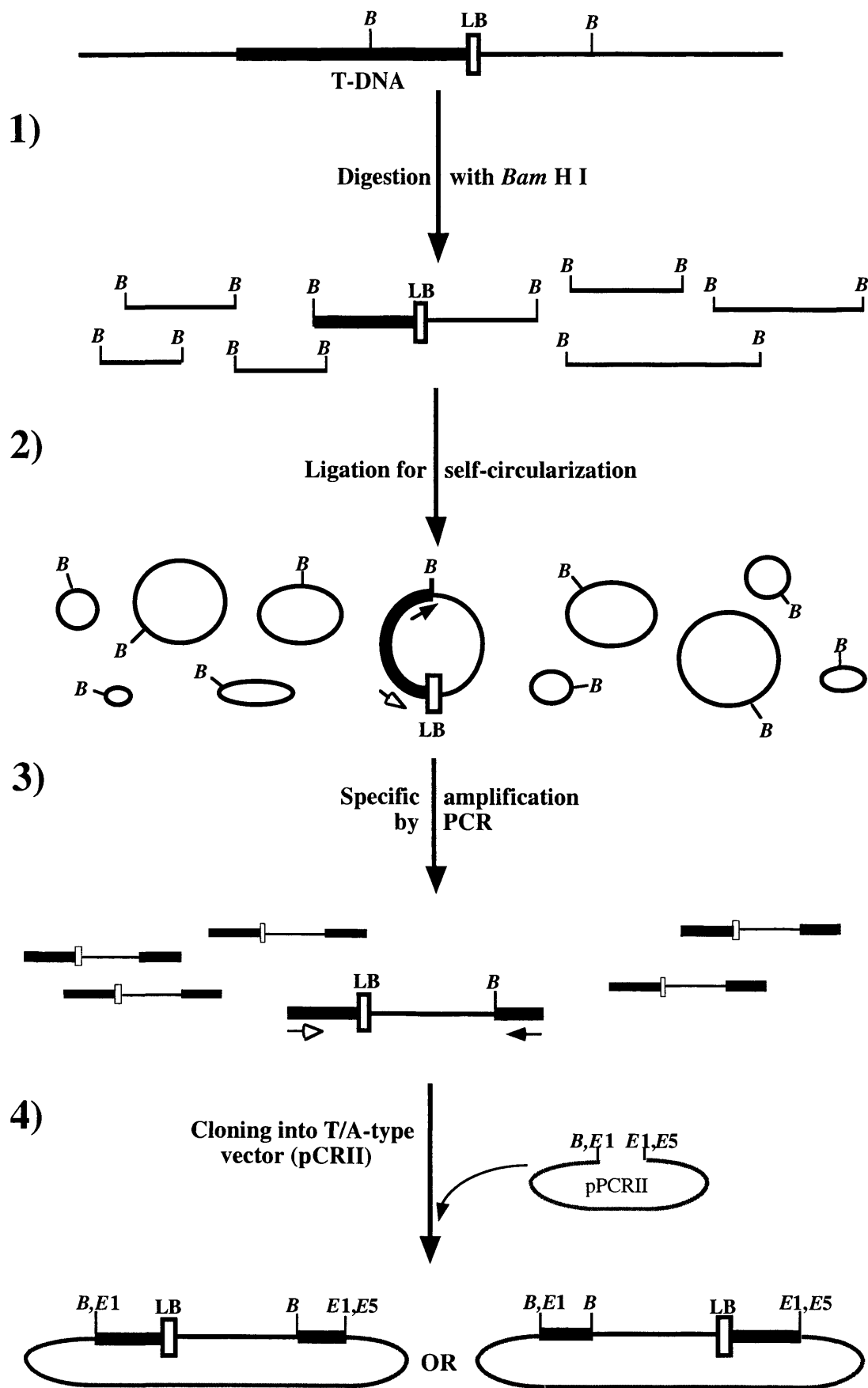
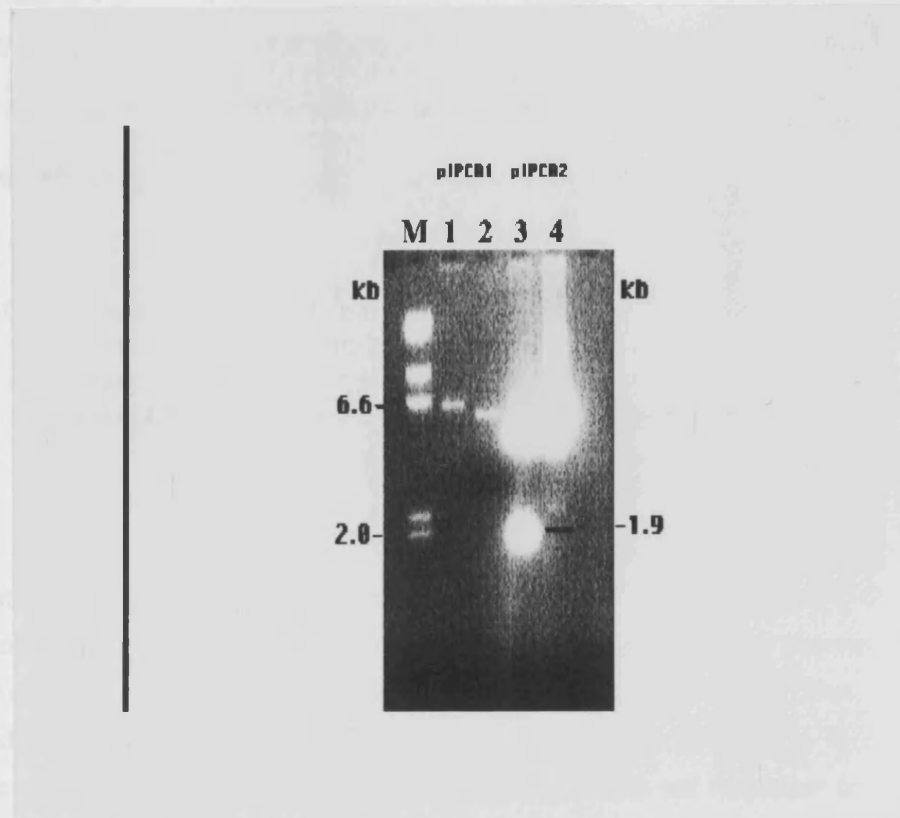


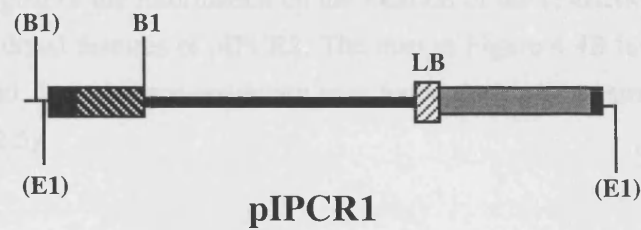
Figure 4.2 Determining the orientation of the inserts in pIPCR1 and pIPCR2.

- A)** Samples of the plasmids pIPCR1 and pIPCR2 were digested with *Bam*H I and resolved by agarose gel electrophoresis together with samples of uncut plasmids. Lane M: λ /*Hind* III DNA marker; Lane 1: uncut pIPCR1; Lane 2: pIPCR1 digested with *Bam*H I; Lane 3: uncut pIPCR2; Lane 4: pIPCR2 digested with *Bam*H I. The picture is overexposed to enable the visualization of the pIPCR1 fragments. The 0.4 kb *Bam*H I fragments generated by the digest of pIPCR1 are not visible in the picture due to the small amount of DNA loaded.
- B)** and **C)** Schematic representations of the orientation of the inserts in pIPCR1 and pIPCR2, respectively, as revealed by the analysis. The expected organization of the diverse T-DNA elements in the absence of sequence rearrangements is shown.
- The restriction enzymes sites are indicated with B1 (*Bam*H I) and E1 (*Eco*R I). Those sites located in the polylinker are within parenthesis.

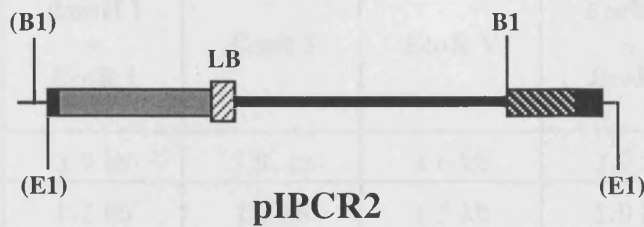
A)



B)



C)



■ M13 sequence ■ *gus* gene sequences — non-T-DNA sequences (native genome)
 ▨ LB sequence ▨ *nos* terminator

pIPCR2

```

.....CTAAGCGTCAATTTGTTTACACCACAATATATTTGAAATTATCTGTAATT.....
      |||||||||||||||||||||||||||||||||||
.....CTAAGCGTCAATTTGTTTACACCACAATATATCCTGCCA

```

LB in p Δ gusBin19

Figure 4.3 Comparison of the T-DNA left border junction in pIPCR2 with the left border (LB) in the transformation vector, p Δ gusBin19 (Topping *et al.*, 1991; Wei, 1994). The two sequences are aligned and matches are indicated by bars. The 7 bp-long terminal segment of T-DNA absent in pIPCR2 is shown in bold italics.

restriction enzyme analysis with *EcoR* I, *EcoR* V, *Hind* III, *Kpn* I, *Pst* I, and *Sph* I. Of these enzymes, only *EcoR* I and *EcoR* V were found to have target sites within the sequence (data not shown). To correctly assign map positions to the *Bam*H I, *EcoR* I and *EcoR* V sites, pIPCR2 was digested with each and combinations of these three restriction enzymes. The digested DNA samples were separated by agarose gel electrophoresis (Fig. 4.4A) and their sizes estimated by comparison with a DNA size marker run in parallel (Table 4.1).

The partial sequences obtained at the T-DNA junction with LB-63 and LB-267 had shown the existence of an *EcoR* V site located 55 bp upstream from the left border. This permitted the merging of the information on the location of the restriction enzyme sites with the expected organizational features of pIPCR2. The map in Figure 4.4B integrates all the available data on the plasmid. Included are restriction sites found during the course of sequencing of the insert (Section 4.2.5).

<i>Bam</i> H I	<i>Bam</i> H I + <i>Eco</i> R I	<i>Eco</i> R I	<i>Eco</i> R V	<i>Eco</i> R V + <i>Eco</i> R I	<i>Bam</i> H I + <i>Eco</i> R V
4.3 kb	3.9 kb	3.9 kb	4.6 kb	3.9 kb	3.9 kb
1.8 kb	1.2 kb	1.2 kb	1.5 kb	1.0 kb	1.1 kb
-	0.6 kb	1.0 kb	-	0.7 kb	0.8 kb
-	0.4 kb	-	-	0.5 kb	0.4 kb

Table 4.1 Sizes of fragments generated by restriction digestion of pIPCR2.

The sizes were estimated from the agarose gel shown in Figure 4.4A by comparison with the fragments of the 1 kb ladder marker (GibcoBRL) run in parallel.

Verification of the general structure of pIPCR2 was performed by partial sequencing of the terminal regions of its insert from the flanking polylinker regions with the forward and reverse universal oligonucleotide primers. The sequences were found to conform to the anticipated organization shown in Figure 4.2C. Also, to confirm that the region upstream from the *Bam*H I site in the insert did not derive from T-DNA, pIPCR2 was digested with *Eco*R V and *Bam*H I. Following agarose gel resolution of the digested DNA, the 1.0 kb *Eco*R V-*Bam*H I fragment was isolated, purified, and then ligated in an orientation-specific manner into pBluescript IITM vector (Stratagene). The vector had been previously linearized with the *Eco*R V and *Bam*H I. The ligated DNA was used to transform *E. coli* as described in Section 2.2.2. After isolation of transformants, their plasmid DNAs were prepared (Section 2.4.1.1). Screening for the appropriate recombinants was performed by digestion of the plasmids with *Bam*H I and separation by electrophoresis in an agarose gel. One of those containing a single insert, designated pIPCR3, was selected and its relevant terminal region manually sequenced. The 248 bp-long partial sequence obtained did not show any resemblance to T-DNA sequences or indeed to any other vector sequences indicating it derived from the native genome of *Arabidopsis*.

4.2.4 Northern blotting hybridization

Examination of the partial sequences obtained for the pIPCR2 insert did not reveal any obvious ORFs (see below, Section 4.2.5). Nevertheless, the relatively long size, about 1 kb, of the non-T-DNA segment (Fig. 4.4B) was thought likely to contain transcribed regions of the wildtype genome in view of the earlier detection of a fusion transcript of about 3.5-4.0 kb (Topping *et al.*, 1994). For this reason total RNA was isolated from *A. thaliana* AtEN101 plants and its wildtype parent cv. C24, as described in Section 2.6.2. The plants utilized were 2 week-old seedlings grown in 1/2 MS10 under continuous illumination at 24 °C. RNA samples were resolved by electrophoresis in a formaldehyde denaturing agarose gel and blotted onto a nylon membrane by capillarity (Section 2.7.1). The plasmid pIPCR3 was digested with *Eco*R V and *Bam*H I, fractionated by agarose gel electrophoresis and its 1.0 kb-long insert fragment was isolated and purified (Section 2.4.6.2). The purified *Bam*H I-*Eco*R V fragment was used to prepare a probe labelled with [α -³²P]dCTP. The Northern blot was hybridized with this probe and washed under standard conditions (Section 2.8.4) but no hybridization was detected even after a relatively long exposure of 3 weeks.

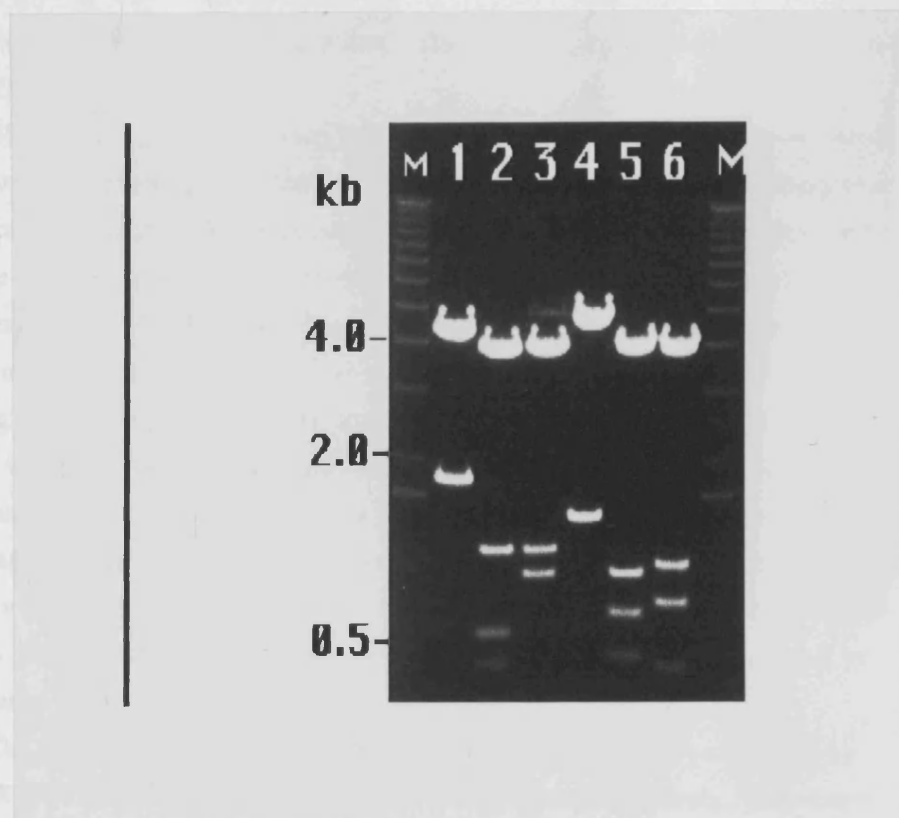
4.2.5 Sequencing of the pIPCR2 insert

There could have been various explanations for the lack of hybridization of the *Eco*R V-*Bam*H I fragment to the C24 and AtEN101 RNAs including a low abundance of the relevant transcripts. Alternatively, the length of transcript target sequence(s) in the IPCR segment might have been insufficient for detection by this means. This could have arisen from the existence in the

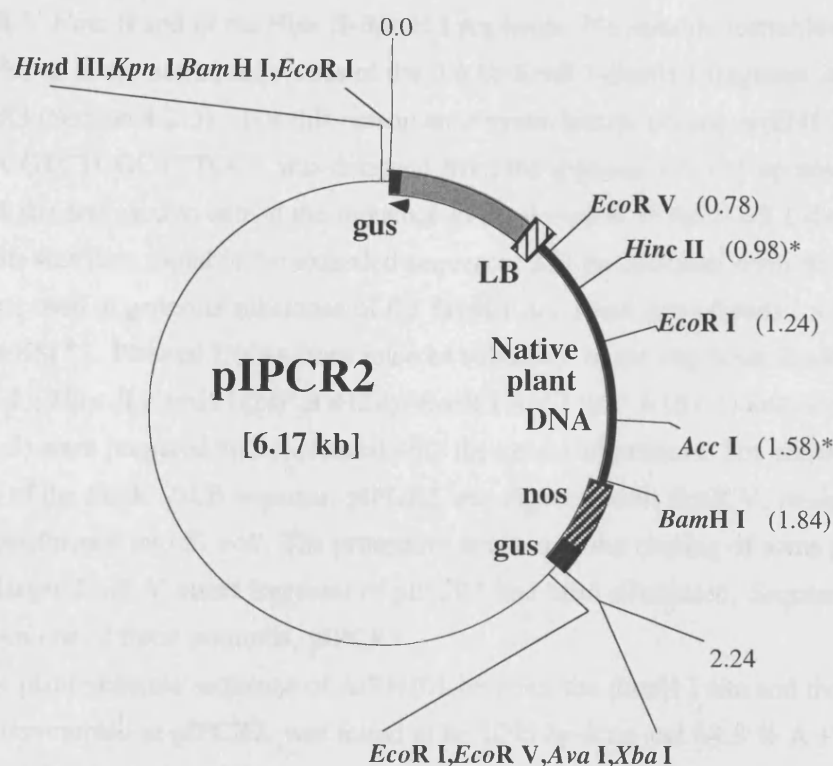
Figure 4.4 Restriction enzyme and features map of insert in pIPCR2.

- A)** Agarose gel of restriction enzyme digests for mapping of sites on plasmid pIPCR2. Lane 1: *Bam*H I; Lane 2: *Bam*H I + *Eco*R I; Lane 3: *Eco*R I; Lane 4: *Eco*R V; Lane 5: *Eco*R V + *Eco*R I; Lane 6: *Bam*H I + *Eco*R V; Lanes M: 1 kb DNA marker ladder.
- B)** Combined restriction enzyme and features map of pIPCR2. A restriction enzyme map was constructed for *Bam*H I, *Eco*R I and *Eco*R V and merged with partial sequence information to develop a map as explained in the text. The map shown includes two sites for *Acc* I and *Hinc* II found during the course of sequencing the insert (labelled with asterisks). The precise positioning of the restriction enzyme sites and the features were derived from the sequence data (Section 4.2.5). Different regions are represented as in Figure 4.2. The arrowhead indicates the direction of transcription of the *gus* gene.

A)



B)



- M13 sequence
- ▨ LB sequence
- *gus* gene sequences
- ▨ *nos* terminator
- non-T-DNA sequences (native genome)

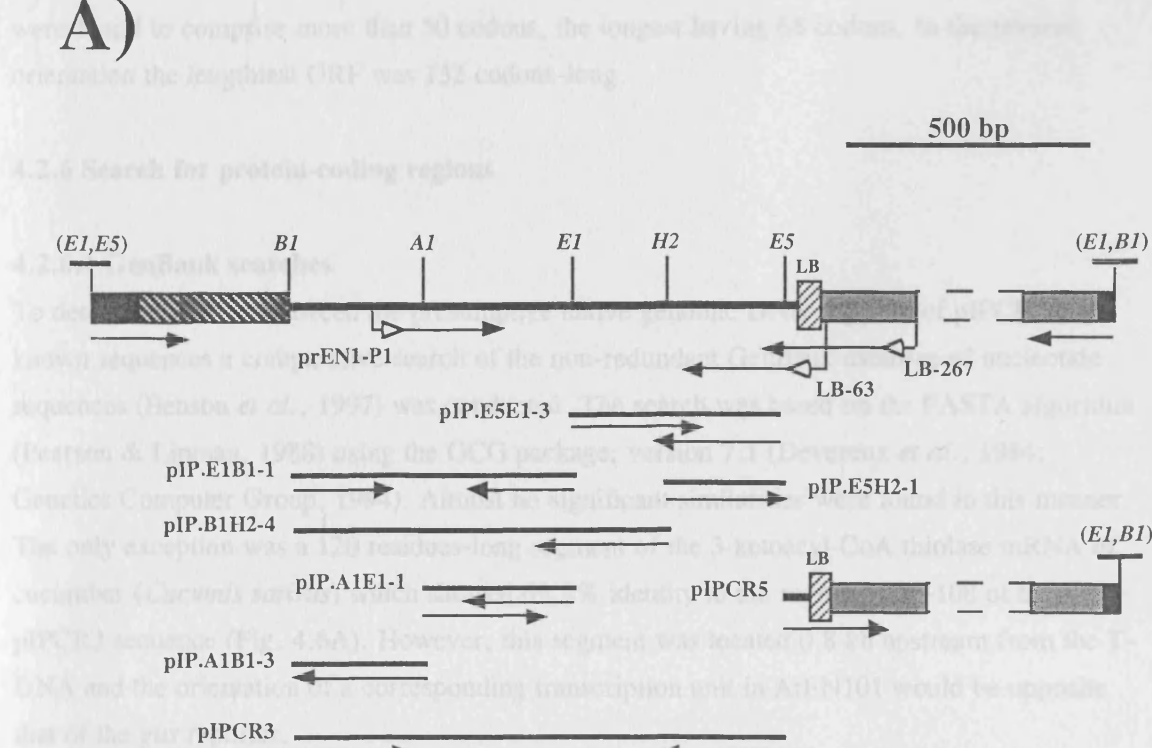
sequence of only short exons. Also, a long intron overlapping the region might have had a similar effect.

The segment located between the T-DNA left border and the upstream *BamH* I site in pIPCR2 was sequenced to gain detailed information on its structure and investigate its capacity for transcript encoding. For sequencing, a number of subclones were generated taking advantage of the known restriction map (cf. Fig. 4.4B). The overall sequencing strategy for pIPCR2 is shown in Figure 4.5A. The plasmid was digested with the appropriate restriction enzymes to isolate the 0.5 kb *EcoR* V-*EcoR* I and the 0.6 kb *EcoR* I-*BamH* I segments. Following separation in agarose gels, the fragments were purified and ligated into pBluescript KS(+)™ vector previously linearized with the appropriate combination of enzymes. After transformation of *E. coli*, suitable clones were selected by preparing plasmid DNA (Section 2.4.1.1) and checking for the presence and size of inserts. The chosen designation for these plasmids was pIP.R1R2-X where "IP" indicated the inverse PCR origin of the cloned fragment, "R1" and "R2" the restriction enzyme sites flanking the inserts, and the numeral "x" referred to the particular isolate.

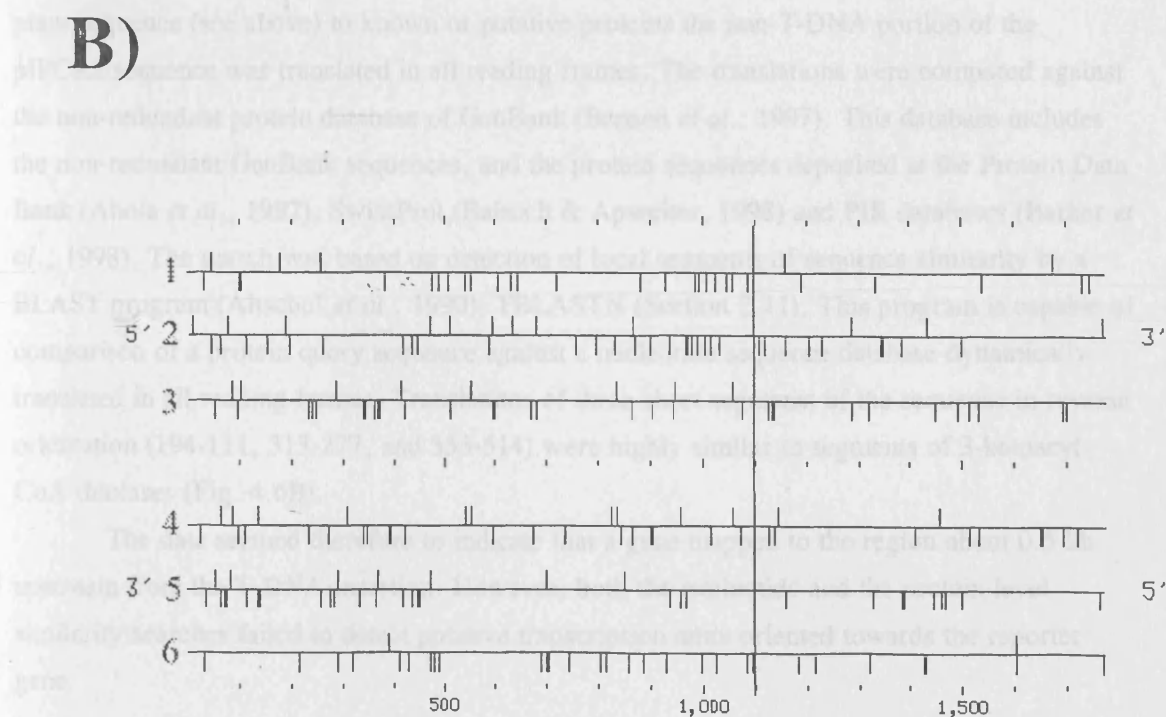
The plasmids pIP.E5E1-3 and pIP.E1B1-1 containing a copy of the 0.5 kb *EcoR* V-*EcoR* I and the 0.6 kb *EcoR* I-*BamH* I fragments, respectively, were sequenced with the universal forward and reverse primers. An *Hinc* II site was found within the *EcoR* V-*EcoR* I fragment, 212 bp upstream from the *EcoR* I site. This was used to generate further subclones of the *EcoR* V-*Hinc* II and of the *Hinc* II-*BamH* I segments. No suitable restriction enzyme sites were found in the partial sequences of the 0.6 kb *EcoR* I-*BamH* I fragment in pIP.E1B1-1 or in pIPCR3 (Section 4.2.3). For this reason an oligonucleotide primer, prEN1-P1 (CCTCCACGTCTCGCTTTGC), was designed from the segment 153-171 bp downstream from the *BamH* I site and used to extend the sequence in the direction of the *EcoR* I site (Fig. 4.5A). An *Acc* I site was then found in the extended sequence, 259 bp upstream from the *BamH* I site. This site was used to generate subclones of the *EcoR* I-*Acc* I and *Acc* I-*BamH* I segments in pBluescript KS(+). Plasmid DNAs from selected subclones of the segments *EcoR* V-*Hinc* II (pIP.E5H2-1), *Hinc* II-*BamH* I (pIP.B1H2-4), *EcoR* I-*Acc* I (pIP.A1E1-1) and *Acc* I-*BamH* I (pIP.A1B1-3) were prepared and sequenced with the universal primers. For second strand sequencing of the *EcoR* V-LB segment, pIPCR2 was digested with *EcoR* V, re-ligated and the products transformed into *E. coli*. The procedure resulted in the cloning of some plasmids where the larger *EcoR* V insert fragment of pIPCR2 had been eliminated. Sequencing was performed on one of these plasmids, pIPCR5.

The plant genomic sequence of AtEN101 between the *BamH* I site and the T-DNA left border, as represented in pIPCR2, was found to be 1095 bp-long and 64.3 % A+T-rich (see below, Fig. 4.8). Analysis of the sequence showed the presence of a total of 127 stop codons in all reading frames (Fig. 4.5B). These effectively limited the size of potential protein-coding segments to less than 40 codons for the most part. Only two ORFs oriented towards the T-DNA

A)



B)



were found to comprise more than 50 codons, the longest having 68 codons. In the reverse orientation the longest ORF was 132 codons-long.

4.2.6 Search for protein-coding regions

4.2.6.1 GenBank searches

To detect similarities between the presumptive native genomic DNA segment of pIPCR2 and known sequences a comparative search of the non-redundant GenBank database of nucleotide sequences (Benson *et al.*, 1997) was conducted. The search was based on the FASTA algorithm (Pearson & Lipman, 1988) using the GCG package, version 7.1 (Devereux *et al.*, 1984; Genetics Computer Group, 1994). Almost no significant similarities were found in this manner. The only exception was a 120 residues-long segment of the 3-ketoacyl-CoA thiolase mRNA of cucumber (*Cucumis sativus*) which showed 69.2% identity to the segment 226-108 of the pIPCR2 sequence (Fig. 4.6A). However, this segment was located 0.8 kb upstream from the T-DNA and the orientation of a corresponding transcription unit in AtEN101 would be opposite that of the *gus* reporter.

Coding regions of genes evolve more rapidly than the cognate protein sequences. As a result, known proteins with similarity to potential protein(s) encoded in the pIPCR2 sequence should be more readily detectable. To search for similarities of the ORFs found in the native plant sequence (see above) to known or putative proteins the non-T-DNA portion of the pIPCR2 sequence was translated in all reading frames. The translations were compared against the non-redundant protein database of GenBank (Benson *et al.*, 1997). This database includes the non-redundant GenBank sequences, and the protein sequences deposited at the Protein Data Bank (Abola *et al.*, 1997), SwissProt (Bairoch & Apweiler, 1998) and PIR databases (Barker *et al.*, 1998). The search was based on detection of local segments of sequence similarity by a BLAST program (Altschul *et al.*, 1990), TBLASTN (Section 2.11). This program is capable of comparison of a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames. Translations of three short segments of the sequence in reverse orientation (194-111, 313-277, and 553-514) were highly similar to segments of 3-ketoacyl-CoA thiolases (Fig. 4.6B).

The data seemed therefore to indicate that a gene mapped to the region about 0.5 kb upstream from the T-DNA insertion. However, both the nucleotide and the protein level similarity searches failed to detect putative transcription units oriented towards the reporter gene.

Figure 4.6 Results of similarity searches for the non-T-DNA sequence in the pIPCR2 insert.

- A)** Edited output of FASTA search of the non-redundant GenBank database (February 1996; Section 2.11).
- B)** Edited output of BLAST (via the program TBLASTN) search of the non-redundant protein GenBank database (February 1996). The reading frames for which significant similarities (random match probability $P < 0.02$) were detected are indicated. The search was performed on-line at NCBI (<http://www.ncbi.nlm.nih.gov/BLAST/>).

A) FASTA results

em_pl:cs3kct /rev X67696 C.sativus mRNA for 3-ketoacyl-C... 142 292 221 Initl Initn Opt

69.2% identity in 120 bp overlap

	220	210	201	191	181	171
pIPCR2	TCTTTGTTTCTTGATTCT-TCTTCTTCTTCAAGGCATATCGTACCGCCATTTGCAAAGCG					
cs3kct	TCGGTGTGGAGATGATGTCGTGATTGTTGCAGCGTACCGTACTGCCATTTGCAAATCG					
	170	180	190	200	210	220

	161	151	141	131	121	111
pIPCR2	AGACGTGGAGGTTTCAAAGACACTCTTCTCTGATGATCTTCTTGCTTCTGTTCTTAAGGTA					
cs3kct	AAGCGTGGTGGCTTCAAGGATACTTATCCCGATGATCTGCTCGCCCCTGTTTGAAGGCA					
	230	240	250	260	270	280

B) BLAST results

Frame 5

				Smallest
				Sum
Sequences producing High-scoring Segment Pairs:	Reading Frame	High Score	Probability P(N)	N
emb X67696 CS3KCT C.sativus mRNA for 3-ketoacyl-CoA thiolase	+2	127	1.6e-08	1
emb X12966 HS3OCOAT Human mRNA for 3-oxoacyl-CoA peroxisomal thiolase	+1	100	0.00011	1
gb J02749 RATKPCOAT Rat peroxisomal 3-ketoacyl-CoA thiolase mRNA, complete cds.	+2	85	0.015	1

emb|X67696|CS3KCT C.sativus mRNA for 3-ketoacyl-CoA thiolase

Query: 280 AYRTAICKARRGGFKDTLPDDLASVLK 307 [in DNA seq: 194-111]

AYRTAICK++RGGFKDT PDDL A VLK

Sbjct: 197 AYRTAICKSKRGGFKDTYPDDLAPVLK 280

emb|X12966|HS3OCOAT Human mRNA for 3-oxoacyl-CoA peroxisomal thiolase

Query: 282 RTAICKARRGGFKDTLPDDLASVL 306

RTAIC+A RGGFKDT PD+LL++V+

Sbjct: 139 RTAICRAGRGGFKDTPDELLSAVM 213

gb|J02749|RATKPCOAT Rat peroxisomal 3-ketoacyl-CoA thiolase mRNA, complete cds.

Query: 282 RTAICKARRGGFKDTLPDDLASVL 306

RT I +A RGGFKDT PD+LL++VL

Sbjct: 158 RTPIGRAGRGGFKDTPDELLSAVL 232

Frame 6

				Smallest
				Sum
Sequences producing High-scoring Segment Pairs:	Reading Frame	High Score	Probability P(N)	N
emb X67696 CS3KCT C.sativus mRNA for 3-ketoacyl-CoA thiolase	+2	72	0.026	2

emb|X67696|CS3KCT C.sativus mRNA for 3-ketoacyl-CoA thiolase

Query: 172 MERAMERQKILLRHLNPVSSS 192 [In DNA seq.: 553-514]

ME+A+ RQ ILL HL P SS+

Sbjct: 24 MEKAINRQSILLHHLRPSSSA 94

Query: 250 AAFGDDIVIVAVY 262 [In DNA seq.: 313-277]

+ FGDD+VIVA Y

Sbjct: 164 SVFGDDVVIVAAY 202

4.2.6.2 Splice sites at the junction region

To further assess the transcript coding potential of the region, the sequence was inspected for possible splice sites. The consensus sequences for 5' splice sites (W₈₃G₇₆/GTW₈₀R₇₂-G₅₂W₇₇W₇₃W₇₇) and 3' splice sites (10xW>₆₅W₈₃R₇₅Y₉₃AG/G₆₅t₄₃) of dicot plants (Simpson *et al.*, 1992) were used for this purpose. Selection of possible 3' splice sites also used, as a criterion, the occurrence of an approximate match to the branch point consensus (Simpson *et al.*, 1992) within 18-40 bp, the known range of their occurrence relative to 3' splice sites in plants (Brown *et al.*, 1996). Few sites oriented towards the T-DNA were identified in this manner (Fig. 4.7) and it was not possible to delimit potential exons in the sequence. The elements found included potential 5' splice sites located at positions 93, 184, 644 and 753, and possible 3' splice sites at positions 121, 144, 584, 998 and 1176. The relative distribution of the various elements (Fig. 4.8) shows that it was conceivable that an exon terminating at any 5' splice site in this region, or located further upstream, could be spliced to a point near the T-DNA left border or beyond it (see below).

4.2.6.3 Potential splice sites in the T-DNA left border region

An inherent structural fault of the binary plasmid pΔ*gus*Bin19 is the large distance (684 bp) separating the T-DNA left border from the initiation codon for the reporter gene *gus*. This resulted from the inadvertent introduction of a segment of the M13 phagemid during the construction of plasmid pBin19 (Bevan, 1984) from which pΔ*gus*Bin19 derives (Topping *et al.*, 1991). Examination of the T-DNA segment upstream from the *gus* gene showed the existence of a total of 25 stop codons in the forward reading frames. Of these, 9 are in-frame with *gus* (Fig. 4.5B, frame 2). Therefore, GUS-fusion proteins can only be produced in the event of splicing out of the stop codons, or by truncation of the LB region during T-DNA integration.

The T-DNA in the segment comprising 750 bp downstream from the LB was inspected for matches to the 3' splice site consensus. One site was found at position -101 that satisfied the criteria of selection (defined in Figure 4.7). Two other sites that nearly satisfied the criteria of selection (I% = 53.8%, cf. Fig. 4.7) were found at positions -276 (TGAGTGTGTTCCAGTT) and -700 (TGTTACGTCCTGTAGAT). The latter occurs downstream from all the stop codons in the frame of the *gus* gene. The site at -700 is placed 14 bp upstream from the initiation codon of *gus*. In AtEN101, use of any of the potential 3' splice sites identified in the T-DNA structure in conjunction with an upstream 5' splice site would result in elimination of sequences represented in pIPCR3 from any GUS-fusion primary transcripts.

Figure 4.7 Potential 5' and 3' splice sites in the T-DNA insertion region upstream from the *gus* gene.

Potential sites were selected by their level of matching to consensus sequences compiled by Simpson *et al.* (1992). For those positions where a base had at least 85% probability of occurrence a perfect match to the consensus was required. For positions with a probability of occurrence of between 60% to 85% an identity percentile was determined (**I%**). An average Conservation percentile was also calculated for the residues in the consensus sequence (**C%**). Potential splice sites were required to have $I\% > C\% - (C\% - m\%)/2$. **m%** refers to the minimal identity percentiles of known splice site sequences identified in a list compiled from *A. thaliana* introns by Simpson *et al.* (1993).

Potential 5' splice sites

Consensus:

W₈₃G₇₆:G₁₀₀T₁₀₀W₈₀R₇₂G₅₂W₇₇W₇₃W₇₇

C%=76.9%

Example of least conserved sequence:

β-tubulin-4 gene (intron 2)	ct:GTgAGTAc	m%=42.9%
Plasma membrane H ⁺ -ATPase gene (intron 9)	ca:GTgAGTNN*	m%=57.1%
Peroxidase gene <i>prxEa</i> (intron 1)	gt:GTAAGgAT	m%=57.1%

Minimum I% > {C%-(C%-m%)/2} = 67%

Pos.	Seq.	I%
91	AGGTTtaATT	85.7%
182	cGGTAcGATA	71.4%
642	TGGTTGaTTT	100.0%
750	TGGTctGATT	71.4%
=====		
-601	AaGTTGGgTA	71.4%

Potential 3' splice sites

Consensus:

W₇₃W₇₇W₇₃W₇₆W₇₆W₆₆W₇₀W₇₃W₇₂W₇₃W₈₃R₇₅Y₉₃A₁₀₀G₁₀₀:G₆₅t₄₃

C%=73.3%

Example of least conserved sequence:

Chlorophyll <i>a/b</i> -binding prot. (intron 2)	TgggTggTgTTTTAGc	m%=38.5%
Histone <i>H1-1flk</i> gene (intron 1)	TcTgggcAggAGCAGa	m%=38.5%

Minimum I% > {C%-(C%-m%)/2} = 55.9%

Branch consensus: **W₇₂t₃₆Y₉₂T₁₀₀R₇₂A₁₀₀T₅₇**

branch	dist.	pos.	seq.	%I
N/A	N/A	13/28	TTcTTTccAcTACAG:c	62.1
GTTTAAT	22	106/121	TTAccTTAAgAACAG:a	69.2
gATTGAg	33	529/144	TTTgTcTTTccATAG:c	62.1
cCaTGAT	30	569/584	AcAcTTgTgTTGCAG:c	62.1
TTTTGAT	17	983/998	ATTTgAgTAATtTAG:a	69.2
AGaTAAT	24	1061/1076	TTgTATTcTTTACAG:a	76.9
=====				
TTTTAAT	28	-101	TTTTcTTTTcAcCAG:t	69.2

Figure 4.8 Partial sequence of the pIPCR2 insert.

Shown are the sequences downstream from the *Bam*H I site up the 5' end of the *gus* gene. Italics indicate T-DNA sequences. Lower case letters represent the non-sequenced segment within the T-DNA left border region (cf. Fig. 4.5) which is presumed intact. GT and AG indicate the potential 5' and 3' splice sites identified in the sequence; possible branch point consensus sequences are overlined (cf. Fig. 4.7). The arrow marks the first residue of the T-DNA. The *gus* start codon is shown in bold and overlined. Numbering of the sequence increases from the *Bam*H I site up to the LB. T-DNA sequences are assigned negative numbers decreasing from the LB. The *Acc* I, *Bam*H I, *Eco*R I, *Eco*R V and *Hinc* II are also indicated.

1 BamH I
 GGATCCAAAGATGTTCTTTCCACTACAGCCTAATCAGAACAAACAATGAAACATAAGAAA
 61 TCATTGATACTACTACTCCATGCATGCACAAGGTTTAATTGATGACTTACCTTAAGAACA
 121 GAAGCAAGAAGATCATCAGGAAGAGTGTCTTTGAAACCTCCACGTCTCGCTTTGCAAATG
 181 GCGGTACGATATGCCTTGAAGAAGAAGAATCAAGAAACAAAGAAAACCTAAAGAAAAGT
 241 GAAAGTGAAGCAAACCTCAATGAGTATACAcc ITTACGCTACAATCACAATGTCATCTCCAAAAG
 301 CAGCCATTGGGGAAACTTCAGAAACACAATTCACAGGCTAAAATTAACAAAAGCACAAGT
 361 GAGTCACATAATAAAACAGAGTAAAGATCATCAGCTTCCACAAACAATTTCAATGTCTA
 421 ATTAAGACTCACGGATTAAGACAACCTAACTTACAGACAGAAGAGAAGGTTTCATGTTTAA
 481 GAGAAGAATTAGAAGAAGAACTGGATTGAGATGACGAAGCAATATCTTTTGTCTTTCCA
 541 TAGCTCTCTCCATGATTCTCTCTGTTGCAAACACTTGTGTTGCAGCTGAATTCTAATGGT EcoR I
 601 GATTGAAAGACAAAACTTTCCGGGGAAACTGAAATGGGTTTGGTTGATTTTTCTGGAAA
 661 GGGGAAGACTTTAAGTACAATGAGCTTCAGGGTCCTCATTTTTATCGTGGGAGAGATTTT
 721 TGATCGATCGTTCTCAGTTCTCACAAGTCTGGTCTGATTGGCCTCGTTGCCAAAACCTTG
 781 TCTTGTCGTATTTGTATAAGTAAEcoR VCGTTCGTTACGAAAATATCGCATTTTTTGACATGTTTT
 841 ACGCCACATACAAATCACGCCGTTAAAAATGTGACCGGGAATTAATCGTCGGCAACTATA
 901 ACTAAACCGGGCCACGTGGGAAGAATCTAACTTTTAAGCAATAAATGTTTCTTCTCATTT
 961 TGGGTATTAGTGATTTTGGATAATTTGAGTAATTTAGAAGAATTTGAGATAATTTATAAAA
 1021 TTTTAGTGAATAATTTGGATATCCAGATAATATACATGATTTGTATTCTTTACAGATAAT EcoR V
 ↓
 1081 TACAGATAATTTCAAATATATTGTGGTGTAACAAATTGACGCTTAGACAACCTTAATAAC
 -46 ACATTGCGGACGTTTTTAATGTACTGGGGTGGTTTTTCTTTTCACCCAGTGAGACGGGCAA
 -106 CAGCTGATTGCCCTTCACCGCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGT
 -166 TTGCCCCAGCAGGCGAAAAATCCTGTTTGATGGTGGTTCGGAATCGGCAAAATCCCTTAT
 -226 AAATCAAAAGAATAGCCCGAGATAGGGTTGAGTGTGTtccagtttgaacaagagtcca
 -286 ctattaaagaacgtggactccaacgtcaaagggcgaaaaaccgtctatcagggcgatggc
 -346 ccactacgtgaaccatcacccaaatcaagtttttggggtcgaggtgccgtaaagcacta
 -406 aatcggaaccctaagggagccccgatttagagcttgacggggAAAGCCGGCGAACGTG
 -466 GCGAGAAAGGAAGGGGAAGAAAGCGAAAGGAGCGGGCGCCATTCAGGCTGCGCAACTGTTG
 -526 GGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGATGTGC
 -586 TGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCAGTCACGACGTTGTAAAACGAC
 -646 GGCCAGTGAATTCCCGGGTGGTCAGTCCCTTATGTTACGTCCTG

4.3 The wildtype genome at the T-DNA insertion site

4.3.1 Rationale

Determining if the expression of a reporter gene reflects that of a wildtype gene requires the identification and study of transcripts mapping to the tagged region in the wildtype genome. For AtEN101, as indicated above (Section 4.2.5), it was very likely that any wildtype mature cognate transcript was either represented in pIPCR2 by short exons, or was altogether absent. Consequently, difficulties were envisaged for the detection of corresponding cDNAs by hybridization with the IPCR clones. It was therefore of interest to acquire information on the structure of the wildtype genome in the region tagged in AtEN101, and then apply this knowledge to obtain probes that might be useful for screening cDNA libraries.

4.3.2 Isolation and characterization of wildtype genomic clones

A genomic library of *A. thaliana* cv. Columbia was screened for clones hybridizing to the insert of pIPCR3 (Section 2.3.2). The library was kindly provided by Dr. J. Dangel (Köln) and had been prepared by Dr. R. Davis (Stanford) in λ GEM 11 (Promega). For screening, a total of about 100,000 plaques were plated. The plaque DNAs were transferred and fixed onto Nylon membranes by established procedures (Section 2.3.2). To prepare the hybridization probe, pIPCR3 was digested with *EcoR* V and *BamH* I and the insert purified after gel fractionation of the digest. The probe was labelled with [α - 32 P]dCTP as described in Section 2.8.2.

Hybridization was at 65 °C and washing of the membranes at the same temperature. Three clones, designated L-EN1-1, -2, and -3, were found to hybridize to the probe and recovered. Two more rounds of plating of the λ isolates, blotting, hybridization with the same probe, and rescue of plaques were necessary for purification of the relevant clones from other contaminating plaques. While this work was in progress, an additional clone, designated L-EN1-4, isolated by others from the same library after hybridization with a probe derived from pIPCR2, was recovered and kindly provided by Dr. Topping.

To confirm the sequence similarity between sequences in the λ clones and the insert of pIPCR3 (*i.e.* the *EcoR* V-*BamH* I fragment, *cf.* Fig. 4.4), DNA from each of the clones was prepared as described in Section 2.4.2.2. Aliquots were digested with *EcoR* I, and, for L-EN1-4, with *BamH* I. The digested samples were electrophoretically separated on an agarose gel (Fig. 4.9A) and blotted onto a nylon membrane. A [α - 32 P]dCTP-labelled probe was prepared from the isolated *EcoR* V-*BamH* I insert as before and hybridized to the Southern blot of the λ clones.

The inserts in λ GEM11 clones are flanked by pairs of *EcoR* I and *BamH* I sites. Thus, digestion of the isolated clones with either *EcoR* I or *BamH* I should release their inserts and the two λ arms of 9 kb and 20 kb. This was observed for all clones except L-EN1-2 where the 9 kb-long λ arm seemed to have been replaced by a 12 kb fragment, presumably as a result of a

mutation (Fig. 4.9A). Internal *EcoR* I sites were also found in the inserts of all the λ clones (Fig. 4.9A; Table 4.2).

The *EcoR* V-*Bam*H I probe was found to hybridize to two *EcoR* I fragments for each clone (Table 4.2; Fig. 4.9), as expected from the restriction map of pIPCR2 insert where an *EcoR* I site is present. The smaller of these fragments (0.5 kb) seemed to be common to all clones. With the available information tentative restriction maps for *EcoR* I sites were derived for each of the clones (Fig. 4.9B). Their construction was based on the observed contiguity of the small λ arm with the presumptive tagged region in L-EN1-2 and the sharing of similar size fragments by some clones. According to the information, the tagged region was most centrally located in the L-EN1-4 clone (Fig. 4.9B). For this reason, this clone was selected for further characterization.

L-EN1-1/ <i>EcoR</i> I	L-EN1-2/ <i>EcoR</i> I	L-EN1-3/ <i>EcoR</i> I	L-EN1-4/ <i>EcoR</i> I	L-EN1-4/ <i>Bam</i> H I
> > 12 kb	> > 12 kb	> > 12 kb	> > 12 kb	> > 12 kb
9 kb	12 kb*	9 kb	9 kb	9 kb
6.2 kb	6.2 kb	6.2 kb	5.7 kb	6.6 kb
4.6 kb*	4.6 kb	4.6 kb	4.6 kb*	5.5 kb*
2.1 kb	0.5 kb*	2.7 kb*	4.3 kb	3.4 kb
0.5 kb*	-	0.5 kb*	0.5 kb*	-

Table 4.2 Restriction fragments obtained by digestion of the genomic λ clones.

The estimated sizes of the restriction fragments visible in the gel shown in Figure 4.9A are indicated. Asterisks label those fragments hybridizing with the probe prepared from the *EcoR* V-*Bam*H I segment of pIPCR2. (Fig. 4.9A).

4.3.3 Restriction map of L-EN1-4

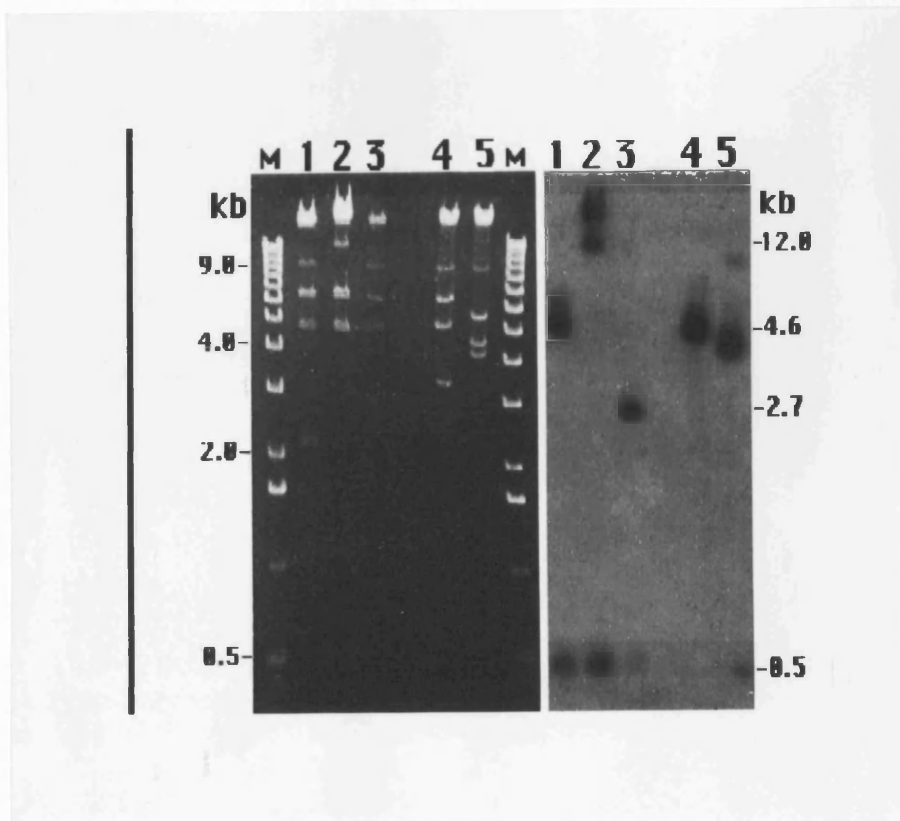
To build a more complete and accurate restriction map for L-EN1-4, its DNA was digested with the enzymes *Bam*H I, *EcoR* I and *Xba* I, and combinations of them. The digests were electrophoresed through agarose gels, an example of which is shown in Figure 4.10A. The sizes of the fragments obtained, estimated by comparison with DNA marker fragments, are listed in Table 4.3. To more accurately assess the sizes of the longer fragments, similar digests were resolved further in agarose gels (*e.g.* Fig. 4.10B).

The fragments of L-EN1-4 with similarity to the insert in pIPCR2 were identified by Southern hybridization. A Southern blot was prepared from the gel in Figure 4.10A. The blot was hybridized to a [α -³²P]dCTP-labelled probe prepared from purified *EcoR* V-*Bam*H I insert of pIPCR3 (Fig. 4.9C). The hybridization was performed under standard conditions (Section

Figure 4.9 Mapping of *EcoR* I sites on wildtype λ genomic clones hybridizing to the region of T-DNA insertion.

- A)** Left: Agarose gel of restriction enzyme digests of the λ clones. Lanes 1-4: *EcoR* I digests of the λ clones L-EN1-1 to -4, respectively; Lane 5: *Bam*H I digest of L-EN1-4. Lanes M: 1 kb DNA marker.
Right: Autoradiographic detection of the fragments with similarity to the T-DNA insertion region. A Southern blot of the gel shown was hybridized to a probe prepared from the purified *Bam*H I-*EcoR* V fragment of pIPCR3 labelled with [α -³²P]dCTP. Hybridization was at 65°C and washes were 2 X 45' Wash I plus 2 X 45' Wash II. Lanes 1-5: as above.
- B)** Proposed *EcoR* I restriction map of the λ genomic clones. The map was constructed based on the *EcoR* I restriction patterns and the hybridization data above. Fragments hybridizing with the probe are shown in grey-filled boxes. The estimated sizes of the various fragments are indicated. The asterisk marking the 3.0 kb segment of L-EN1-2 indicates the absence of an *EcoR* I site separating the genomic DNA from the 9 kb λ arm due to mutation. The proposed size of the genomic portion (3.0 kb) is the estimated extension assuming that the arm is mostly intact.

A)



B)

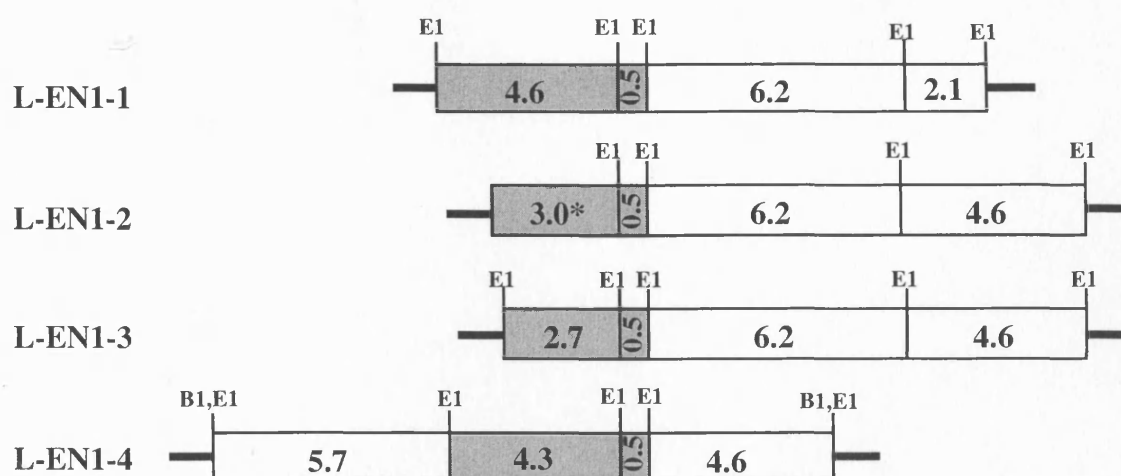
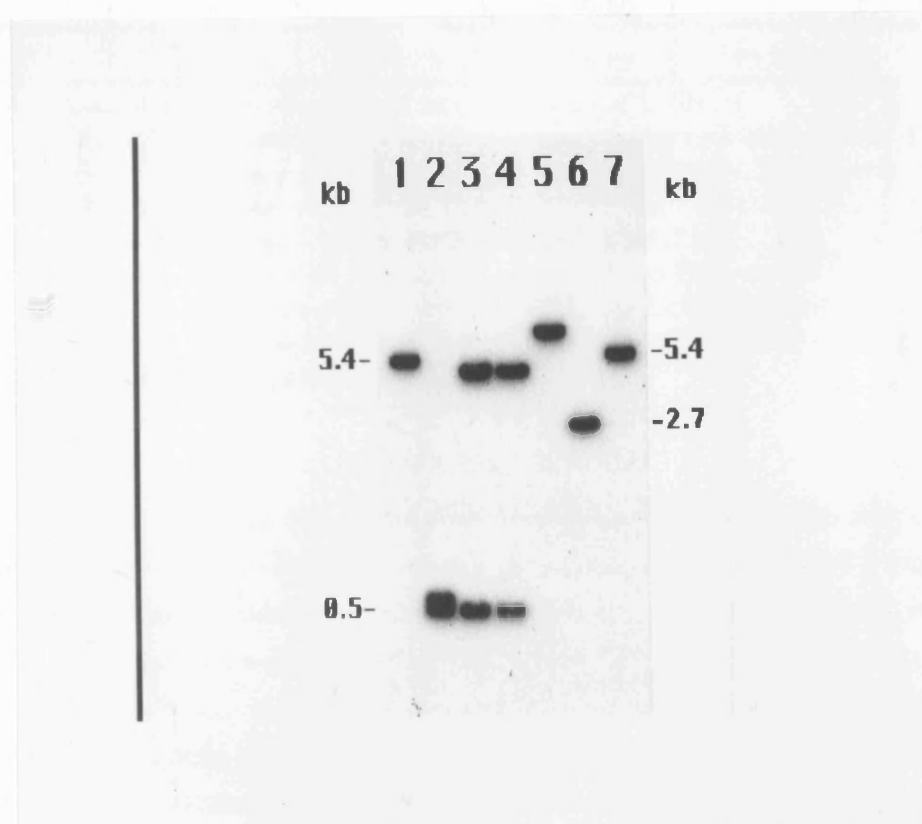
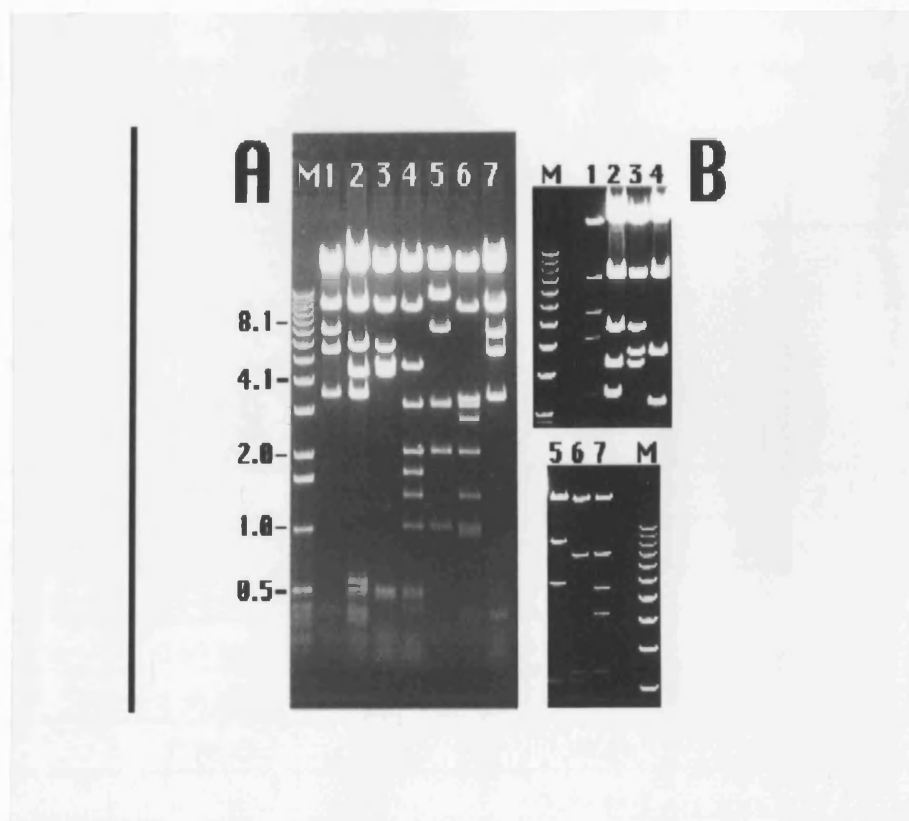


Figure 4.10 Restriction enzyme analysis of L-EN1-4 and identification of T-DNA insertion region.

Top: A) Agarose gel resolution of restriction enzyme digests of the λ genomic clone L-EN1-4 DNA. Lane M: 1 kb DNA marker. Digests: Lanes 1 and 7: *Bam*H I; Lane 2: *Bam*H I + *Eco*R I; Lane 3: *Eco*R I; Lane 4: *Eco*R I + *Xba* I; Lane 5: *Xba* I; Lane 6: *Bam*H I + *Xba* I; Lane 7: *Bam*H I.

B) Example of a longer electrophoretic separation of L-EN1-4 digests identical to those shown in A). Only fragments of at least 3 kb are visible as the others were run out of the gel. Lanes: as in A).

Bottom: Identification of fragments with similarity to the T-DNA insertion region in pIPCR2. A Southern blot of the gel in A) was hybridized to a [α - 32 P]dCTP-labelled probe of the *Bam*H I-*Eco*R V insert of pIPCR3. Hybridization was at 65°C and washes were 2 X 45' Wash I plus 2 X 45' Wash II. Lanes 1-7 as in A).



2.8.3). Together, the restriction and hybridization data allowed the derivation of a restriction map of L-EN1-4 identifying the region with sequence similarity to the tagged region in AtEN101.

<i>Bam</i> H I	<i>Bam</i> H I + <i>Eco</i> R I	<i>Eco</i> R I	<i>Eco</i> R I + <i>Xba</i> I	<i>Xba</i> I	<i>Xba</i> I + <i>Bam</i> H I
6.5 kb	5.7 kb	5.7 kb	4.6 kb #	7.0 kb #	3.4 kb
5.4 kb #	4.3 kb	4.6 kb #	3.2 kb	3.2 kb	3.2 kb
3.4 kb	3.4 kb	4.3 kb	2.1 kb	2.1 kb	2.8 kb #
0.39 kb	0.60 kb #	0.50 kb #	1.7 kb	1.5 kb*	2.1 kb
-	0.50 kb #	0.50 kb	1.5 kb	1.0 kb	1.5 kb-1
-	0.50 kb	-	1.05 kb	1.0 kb	1.0 kb
-	0.39 kb	-	0.50 kb #	-	1.0 kb
-	0.32 kb	-	0.50 kb	-	0.40 kb
-	-	-	0.40 kb	-	0.40 kb
15.7 kb	15.8 kb	15.7 kb	15.6 kb	15.8 kb	15.8 kb

Table 4.3 Restriction fragments obtained from the insert of L-EN1-4.

The fragments listed were identified and their sizes estimated directly from agarose gels of restriction digests of the λ clone and/or, indirectly, from subclones (see text). Whereas digestion with *Bam*H I or *Eco*R I release the insert, only one *Xba* I site flanks it in the polylinker of 20 kb arm of λ GEM11. Thus the 1.5 kb *Xba* I fragment, marked with an asterisk, was identified from the shift of the 9.0 kb arm band and the double digests with *Bam*H I and *Eco*R I. Labelled with "#" are the fragments found to hybridize with the [α -³²P]dCTP-labelled probe prepared from purified *Eco*R V-*Bam*H I insert of pIPCR3.

An improved version of the restriction enzyme map is shown in Figure 4.11. Its increased accuracy was provided by subcloning the various *Bam*H I and *Eco*R I fragments listed and performing similar restriction enzyme site analyses and partial sequencing on several of the subclones. The subclones were labelled pENRx, where "R" refers to the restriction enzyme *Bam*H I (B) or *Eco*R I (E) used for cloning, and "x" indicates the isolate number. For convenience, the various restriction fragments of L-EN1-4 will be named in what follows according to the restriction enzyme sites that delimited them, as defined in Figure 4.11. Two strategies were adopted for subcloning. Firstly, *Bam*H I and *Eco*R I digests of L-EN1-4 and of the cloning vector pUC18 were mixed and ligated together. *E. coli* was transformed with the

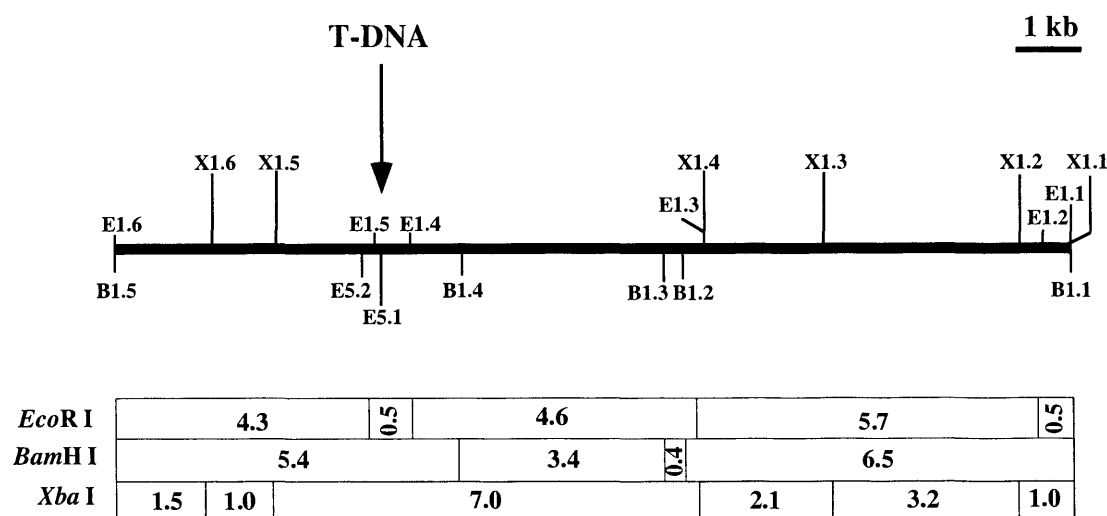


Figure 4.11 Restriction enzyme map of the L-EN-1-4 insert.

Top: Restriction enzyme sites for *BamH* I (B1.x), *EcoR* I (E1.x) and *Xba* I (E1.x) within the insert (represented by the horizontal black bar) and flanking it at the polylinker sites are shown. Below, the fragments generated by digestion with each of these enzymes are represented by the open rectangles where the numbers indicate the sizes in kb. The two *EcoR* V (E5.1 and E5.2) sites flanking E1.5 are also shown. The point of T-DNA insertion in AtEN101 is indicated by the arrow.

ligation mixtures. Transformants were screened for their cloned fragments and appropriate isolates were selected. This approach was successful in the cloning of most *EcoR* I and *BamH* I fragments. Exceptions were the larger B1.1B1.2 (6.5 kb) *BamH* I fragment and the E1.2E1.3 (5.7 kb) *EcoR* I fragment. For these, a direct cloning approach had to be followed involving their purification after gel fractionation and isolation.

Restriction analyses of the various types of pENBs and pENEs were performed by digesting them with either *BamH* I, *EcoR* I or *Xba* I, and combinations of these (*e.g.* Figure 4.12). The usefulness of these analyses is illustrated by the confirmation of the presence of the 0.5 kb E1.1E1.2 fragment within B1.1B1.2 (in plasmid pENB13), distinct from the 0.5 kb E1.4E1.5 fragment, as can be seen in Figure 4.12A (lane 2). Other improvements on the map were provided by sequence analysis. For example, terminal sequencing of the E1.2E1.3 fragment in pENE10 confirmed the presence of a *Xba* I site (X1.4) only 17 residues from the *EcoR* I site.

Some slight disparities were found in the total lengths of the L-EN1-4 insert estimated from each digest (Table 4.3). These can be partially explained by the inaccuracy in determining the migration distances of the larger fragments. Also, the inability to detect smaller fragments by electrophoresis is likely to have contributed for the observed errors. One such fragment was found by terminal sequencing of pENE10 where a small 93 bp *EcoR* I fragment was found in addition to the 4.6 kb E1.3E1.4 fragment.

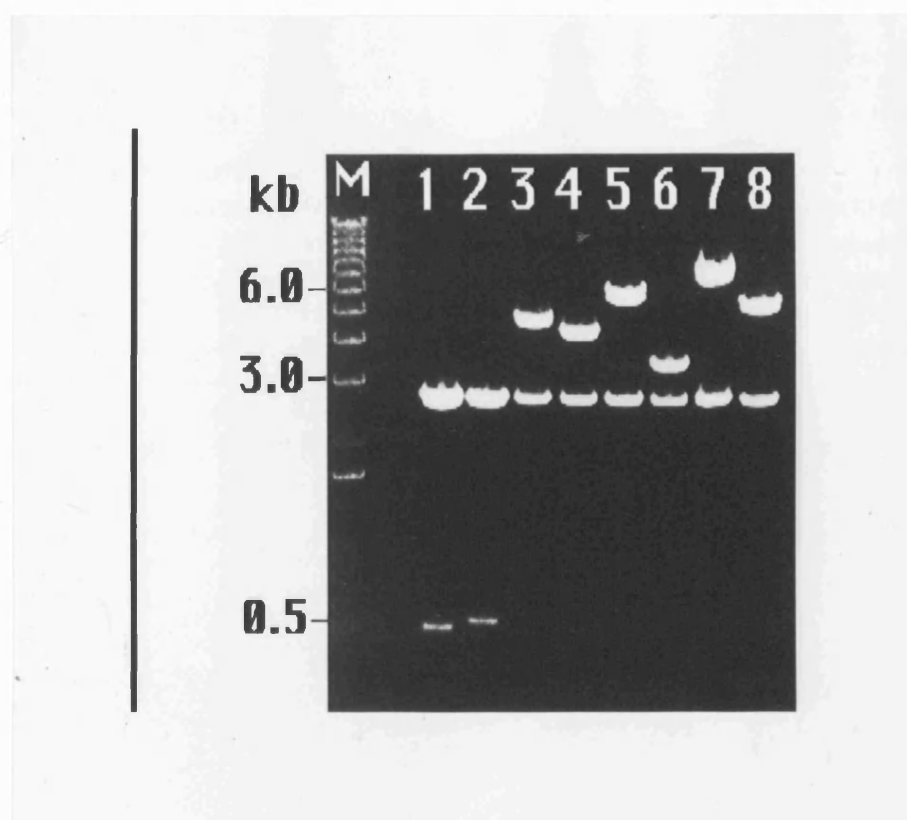
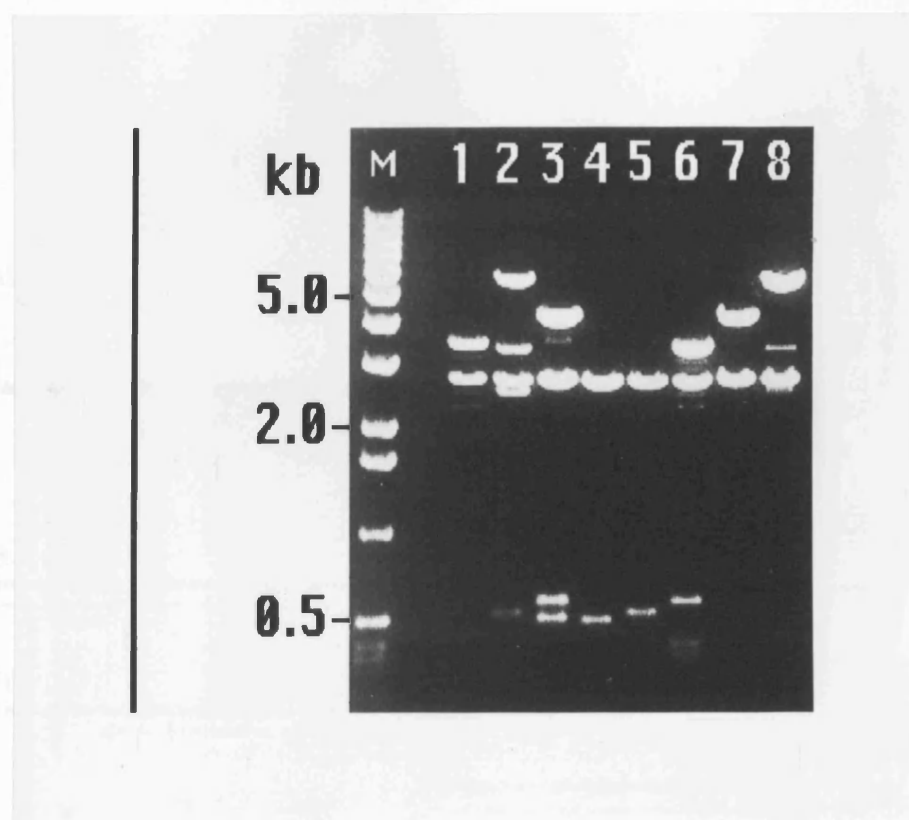
4.3.4 Sequencing the wildtype genome

Several inserts of subclones of L-EN1-4 were partially sequenced. These included the E1.1E1.2 (pENB13), E1.2E1.3, E1.3E1.4 (pENE10), E1.4E1.5 (pENE2), B1.3B1.4 (pENB3) and B1.4B1.5 (pENB8) fragments. Some of the partial sequences were found to be nearly identical to portions of the native plant DNA in the IPCR clones. The complete sequence of the segment between the sites B1.4 and E1.5 was determined. The sequence strategy followed is shown in Figure 4.13. It included the overlapping partial sequences of B1.4B1.5 and E1.3E1.4 in plasmids pENB8 and pENE10, respectively, with the universal reverse primer. The remainder was determined by the complete sequencing of the E1.4E1.5 segment in pENE2 with both directions with the universal primers. The segment was found to be 1092 bp-long and nearly identical to that of the corresponding portion of the IPCR sequence (see Appendix 1). This permitted the tentative joining of the sequences of the B1.4E1.4 and E1.4E1.5 at the E1.4 site. Later sequencing of B1.4B1.5 with primer pREN1-P2 (Section 5.4.2.2) confirmed that B1.4E1.4 and E1.4E1.5 were adjacent. Second strand sequencing of the 96 bp near the B1.4 site was thought not necessary as the first strand sequence was identical to that of the pIPCR2 in the region. The few differences found between the two sequences included two single base mismatches, one single base deletion and an 8 bp insertion in the wildtype sequence relative to that of the IPCR product. In addition, at the site of T-DNA insertion, *i.e.* adjacent to the

Figure 4.12 Restriction enzyme analysis of several subclones of L-EN1-4.

Top: Combined *Bam*H I and *Eco*R I digests. Lane 1: pENB3; Lane 2: pENB13; Lane 3: pENB13; Lane 4: pENB14; Lane 4: pENE2; Lane 5: pENE3; Lane 6: pENE10; Lane 7: pENE14; Lane 8: pENE20.

Bottom: Sizing the inserts. *Eco*R I digests of pENE clones-Lane 1: pENE2; Lane 2: pENE3; Lane 3: pENE10; Lane 4: pENE14; Lane 5: pENE20. *Bam*H I digests of pENB clones-Lane 6: pENB3; Lane 7: pENB13; Lane 8: pENB14.



Genomic fragment

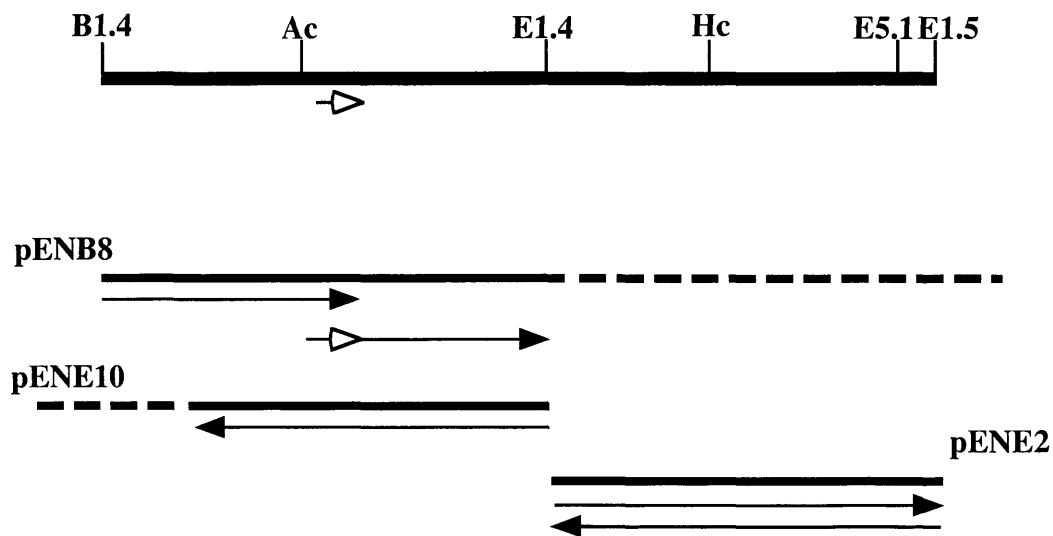


Figure 4.13 Strategy for sequencing the B1.4E1.5 segment of L-EN1-4.

Automatic sequencing was performed on plasmids pENB8, pENE10 and pENE2 with the universal primers. The identity of the sequences on each side of the E1.4 site to the equivalent region in pIPCR2 indicated that they were adjacent. This was later confirmed by sequencing pENB8 with prEN1P-2. The genomic and plasmid fragments are represented as thick bars. Non-sequenced portions of the plasmids are shown as discontinuous lines (not to scale). Arrows indicate the direction and length of the determined sequences. The location of the prEN1-P2 sequence is marked by the white arrowhead

starting point of similarity to the T-DNA, the IPCR and the wildtype sequences are mismatched, as can be seen in Figure 4.14. These alterations in the B1.4E1.5 sequence resulted in the loss of the E1.5 site in AtEN101.

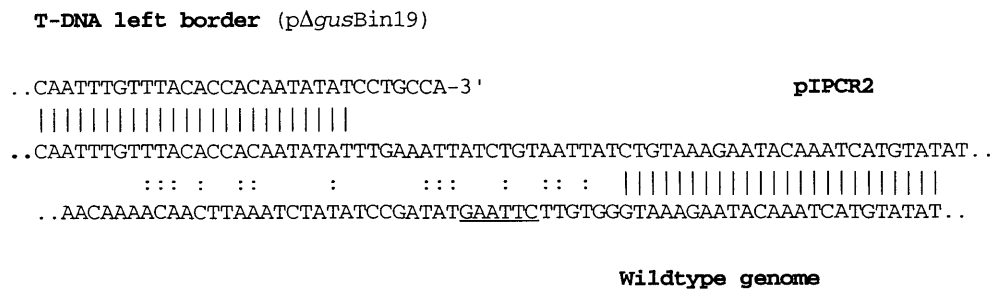


Figure 4.14 Derivation of the junction region at the T-DNA left border in AtEN101.

The junction segment in the pIPCR2 insert is shown aligned with the wildtype genomic sequence and the LB of the transformation vector p*AgusBin19*. Identity of bases resulting from homology is indicated by the vertical bars "|". The *EcoR* I site absent in the IPCR product is underlined in the wildtype sequence.

4.4 Isolation of cDNAs mapping to the T-DNA tagged region

For the identification of wildtype transcripts mapping to the T-DNA tagged region an unusual strategy for screening a cDNA library was followed. It consisted of probing the library with two relatively large genomic fragments extending in opposite directions but overlapping in the region of interest. This strategy had a dual purpose:

- i) Identification of transcription units extending over the T-DNA insertion point, allowing for its detection even if the latter is located close to a large intron.
- ii) The identification of other genes located in the vicinity with a view for later study of possible links of their expressions.

The two fragments of L-EN1-4 selected for the screen were the 5.4 kb B1.4B1.5 and the 4.6 kb E1.3E1.4 (cf. Fig. 4.11). The two fragments were isolated from pENB8 and pENE1, respectively, and used to obtain [α -³²P]dCTP-labelled probes.

The cDNA library screened, generously provided by Dr. J. Giraudat (Gif-sur-Yvette), had been prepared from green silique mRNA (Giraudat *et al.*, 1992). It was based on the λ Zap

II vector (Stratagene) where the cDNA inserts were flanked by *EcoR* I site-bearing adapters. A total of about 200,000 clones from this library were plated. The plaque DNAs on each plate were transferred and fixed onto triple blot filters (Section 2.3.2). Each probe was hybridized to one set of the filters and both probes were concurrently hybridized to the third set of membranes (Section 2.8.3). The probe derived from the 4.6 kb *EcoR* I fragment was found to hybridize with 10 clones, designated pCEN1-1 to -10. Of these only two clones, pCEN1-5 and -8, hybridized also with the probe prepared from the 5.4 kb *BamH* I fragment. The latter probe was not found to hybridize to any other clones. The pCEN-x clones were isolated and subjected to two rounds of purification to eliminate contaminants. During this secondary screening steps it became evident that two clones (pCEN1-3 and -6) were false positives and only the remaining eight were used for further studies.

The λ ZAPII vector contains the necessary elements to allow the *in vivo* retrieval of a pBluescript SK(-) phagemid containing the cloned insert. The excision of the phagemid from the λ clone is dependent on an f1 or M13 helper phage. Phagemids were rescued from the eight λ clones selected by co-cultivation with Ex-Assist helper phage, as described in Section 2.3.3, and their DNA isolated. To confirm the sequence similarity of their inserts to the T-DNA tagged region and to size them, the phagemids were digested with *EcoR* I and a Southern blot of the electrophoretic separation of the digests was prepared. As can be seen in Figure 4.15A, five of the clones originated a single insert whereas two insert fragments were found in pCEN1-5 and -10. The estimated sizes of the inserts are listed in Table 4.4. One of the plasmids, pCEN1-4, was simply linearized by *EcoR* I, presumably due to a mutation affecting one of its *EcoR* I sites at the insert/vector junctions. Assuming that the size of the vector remained unaltered in this clone, the insert should be about 1.0 kb-long.

pCEN1-1	pCEN1-2	pCEN1-7	pCEN1-4	pCEN1-5	pCEN1-8	pCEN1-9	pCEN1-10
1.5 kb	1.1 kb	1.0 kb	4.9 kb*	1.5 kb	1.5 kb	1.6 kb	1.2 kb
-	-	-	-	0.6 kb	-	-	0.4 kb

Table 4.4 *EcoR* I insert fragments in the pCEN1-x cDNA clones.

Listed are the insert-derived restriction fragments visible in the gel shown in Figure 4.15. The asterisk marks the only fragment originated by digestion of pCEN1-4 and which includes the vector sequences.

Whereas the hybridization of pCEN1-5 and -8 to both probes indicated that their cDNAs should map close to or overlap the T-DNA insertion point, it was unclear where the remainder cDNAs mapped to. For this reason, the blot was probed with the 2.2 kb *Pst* I-*EcoR* I segment at the proximal end of the 4.6 kb *EcoR* I fragment relative to the T-DNA insertion point (cf. map, Fig. 4.11). As can be seen in Figure 4.15B, the inserts of seven clones were

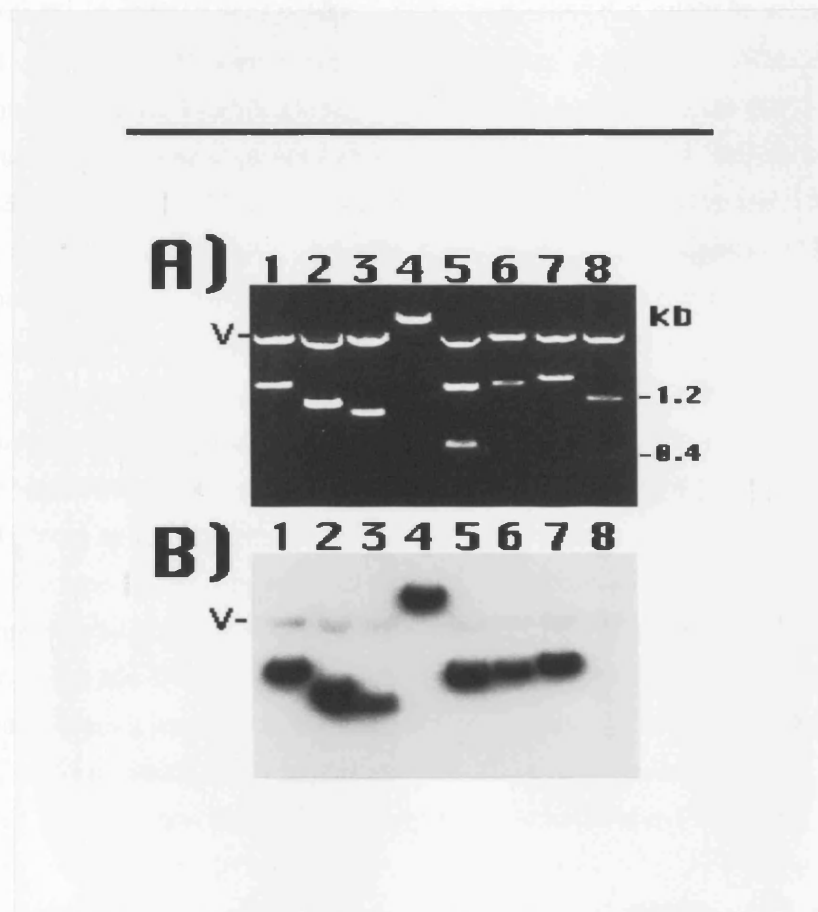


Figure 4.15 Sizing the putative cDNAs and confirming their similarity to the T-DNA insertion region.

- A)** Agarose gel separation of the products of *EcoR* I digestion of DNA of the selected cDNA clones. Lane 1: pCEN1-1; Lane 2: pCEN1-2; Lane 3: pCEN1-7; Lane 4: pCEN1-4; Lane 5: pCEN1-5; Lane 6: pCEN1-8; Lane 7: pCEN1-9; Lane 8: pCEN1-10.
- B)** Autoradiography of a Southern blot of the gel in A) hybridized to a probe prepared from the 2.2 kb *Pst* I-*EcoR* I segment at the proximal end of the 4.6 kb *EcoR* I fragment (E1.3E1.4) relative to the T-DNA insertion point. The probe was prepared by the random oligonucleotide priming method and labelled with [α - 32 P]dCTP. Hybridization and washes were conducted at 65°C.

found to hybridize to this probe, with pCEN1-10 being the exception. Also, for pCEN1-5, the probe hybridized only to the 1.5 kb *EcoR* I fragment.

4.5 Summary

Efforts towards the identification of wildtype native transcripts that might be related to the expression of *gus* in AtEN101 were reported in this chapter. A previously cloned IPCR product of the junction between the T-DNA LB region and the native plant genome (Dr. Topping, personal communication) was sequenced and used to screen a Northern blot of total RNAs of wildtype *Arabidopsis* and AtEN101 (Section 4.2). The Northern hybridization failed to detect any wildtype transcripts and analysis of the IPCR product sequence (in plasmid pIPCR2) did not reveal potential genes in the region with the same orientation as that expected for the reporter gene in AtEN101. The above results and the characteristics of the sequence in the IPCR product suggested that longer segments of the genome were required for detection of transcripts mapping to the region. Specifically, it was possible that any wildtype transcript related to the fusion transcript previously detected in AtEN101 (Topping *et al.*, 1994) was poorly (short exons) or not at all represented in pIPCR2.

As a first step in obtaining suitable probes for detection of transcripts, a λ bacteriophage wildtype genomic library was screened with the insert of pIPCR2 for isolation of clones encompassing the T-DNA insertion site (Section 4.3). Four such clones were isolated and a detailed restriction map was constructed for one of them, L-EN1-4. The *Bam*H I-*Eco*R I segment of this clone, which contained the site of T-DNA insertion in AtEN101, was completely sequenced. It was found to be nearly identical to the non-T-DNA segment of the pIPCR2 insert.

Two overlapping segments of the wildtype genome, spanning about 9.6 kb and including the T-DNA insertion site, were screened for transcription units represented in a cDNA library (Section 4.4). Eight cDNA clones were found to hybridize to one or both genomic segments. The analysis of the corresponding transcription units are the subject of the next chapter.

Chapter 5

Two overlapping genes in the tagged region

5.1 Introduction

The experiments described in Section 4.4 resulted in the identification of cDNAs with sequence similarity to the wildtype genomic region near the T-DNA insertion point. However, the observed similarity was no proof of identity of the sequences. This issue arises with particular acuteness in relation to the hybridization of molecules that are likely to differ by the presence of intervening sequences. It was then necessary to determine if the similarities detected reflected the existence of a transcription unit in the region. Alternatively, the putative transcripts might map elsewhere in the genome. The nature of the clones and their genomic correspondence were firstly investigated by sequence analysis, as reported in this chapter (Sections 5.2 and 5.3). The sequencing data established that the clones contained *bona fide* cDNA inserts. These cDNAs were found to correspond to two different types of transcript. Comparative analysis of the cDNA sequences confirmed that they, in all likelihood, originated from the genomic region of T-DNA insertion. The two transcript types were identical for most of their length. The region of identity started internally, their first exons being dissimilar. As described in Chapter 7, long ORFs were found in both cDNA types and the corresponding genes were identified as 3-ketoacyl-CoA thiolases and designated *PKT1* (pCEN1-8 type) and *PKT2* (pCEN1-9 type). The intron separating the first two exons of *PKT1* represented in pCEN1-8 was about 1 kb-long and contained the first exon of pCEN1-9 (*PKT2*) as well as the site of T-DNA insertion.

Unexpectedly, the orientation of the *PKT1/PKT2* genes was opposite to that of the reporter gene. Their proximity to the T-DNA insertion point suggested, however, that there might be a link between the promoter of *PKT2* and the expression of *gus* in AtEN101. In Section 5.4 is described how, to investigate this possibility, the known 5' end region of the *PKT2* transcript was extended and cloned in order to localize its associated promoter in the genome. The 5' terminal region of the *PKT1* transcription unit was similarly investigated. With the same objective, the database of *A. thaliana* Expressed Sequenced Tags (ESTs; Flanders, 1998) was searched for possible cDNAs of *PKT1* and *PKT2* transcripts (Section 5.5). One EST clone was found to correspond to *PKT1*.

5.2 Sequencing the cDNAs

5.2.1 Partial sequencing reveals similarity to the tagged genome

Preliminary information on the structure of the cDNAs was obtained by partial sequencing of their terminal regions. Unidirectional sequencing was performed manually by the dideoxy method using the universal forward and reverse primers (Section 2.5.4.1). With the exception of pCEN1-4 and pCEN1-5, all clones displayed poly(dA) tracks longer than 15 nt at one of their termini indicating that they contained genuine cDNAs. It was not possible to obtain a sequence for one of the termini of the pCEN1-4 insert presumably because the universal reverse primer binding site had been altered in this clone (cf. Section 4.4). Five of the cDNAs, in clones pCEN1-1, -4, -5, -8, and -9, shared most of the sequences at their putative 5' terminal regions, *i.e.* those opposite the terminal poly(dA) tracks of pCEN1-1, -8 and -9. They were also found to be identical to segments of the genomic sequence between B1.4 and E1.5 (cf. Fig. 4.11). These genomic segments were separated by intervening sequences (introns), further evidence that the clones, including pCEN1-4 and pCEN1-5, did correspond to cDNAs (Fig. 5.1). Unexpectedly, the putative transcripts were oriented in the opposite direction of that deduced for *gus* from the IPCR product.

The five clones, pCEN1-1, -4, -5, -8 and -9, fell into two groups according to the sequences at the 5' ends of their cDNAs. For three phagemids, pCEN1-1, pCEN1-4 and pCEN1-9, the 5' ends mapped within the same segment (exon) of the B1.4-E1.5 genomic region (Fig. 5.1). The 5' terminal sequence of pCEN1-8 was essentially identical to one of the termini of pCEN1-5, only shorter by 7 bp. The identity of these two cDNAs to the genomic sequence of B1.4E1.5 did not begin at their 5' ends (Fig. 5.1). Instead, it started internally coinciding with the start of the second exon of pCEN1-9. This observation could be explained by either of two alternative arrangements: the corresponding transcripts might originate at a considerable distance upstream (> 759 bp), past the point of T-DNA insertion; Or, they could result from the existence of a gene homologous to that represented in pCEN1-9 mapping elsewhere in the genome.

For those cDNAs for which it was possible to determine the 3' end sequences, two groups were distinguished. Although essentially identical, the two groups differed in the position of their poly(dA) tails. In two clones, pCEN1-1 and pCEN1-9, the cDNAs shared the sequences immediately upstream from their polyadenylation sites. For the four other clones, pCEN1-2, -7, -8 and -10, polyadenylation had occurred 21 bp further upstream.

5.2.2 The cDNAs correspond to two distinct but highly related transcripts

The failure to detect other transcripts in the region with the same orientation as *gus* raised the possibility that the expression of the reporter in the transgenic line might be related in some manner to the transcription units typified in the pCEN-x clones. This made it of interest to

Figure 5.1 Alignment of the 5' ends of the cDNAs with the genomic fragment B1.4E1.5. Intron sequences absent in the cDNAs are shown as dashes. Segments of some introns are abbreviated and the length of the missing sequences indicated. In the genomic sequence matches to the conserved dinucleotides GT and AG at the 5' and 3' splice sites are underlined. Numbering of the B1.4E1.5 sequence begins at the *Bam*H I site. The B1.4 and the E1.4 sites are labelled and shown in underlined bold typeface.

pCEN1-9 CCCATTTTCAGTTTCCCCGAAAGTTTGTCTTTCAATCACCATTAGAATTCAGCTGCAA 60
 B1.4E1.5 CCCATTTTCAGTTTCCCCGAAAGTTTGTCTTTCAATCACCATTAGAATTCAGCTGCAA 580

EcoR I

pCEN1-4 TGC 3
 pCEN1-9 CACAAGTGTTTGCAACAGAGAGAATCATGGAGAGAGCTATGGAAAGACAAAAGATATTGC 120
 B1.4E1.5 CACAAGTGTTTGCAACAGAGAGAATCATGGAGAGAGCTATGGAAAGACAAAAGATATTGC 520

pCEN1-1 CCAGTTTCTTCTTCTAATTCTTCTCTTAAACATGAACCTTCTCTTC 46
 pCEN1-4 TTCGTCATCTCAATCCAGTTTCTTCTTCTAATTCTTCTCTTAAACATGAACCTTCTCTTC 63
 pCEN1-9 TTCGTCATCTCAATCCAGTTTCTTCTTCTAATTCTTCTCTTAAACATGAACCTTCTCTTC 180
 B1.4E1.5 TTCGTCATCTCAATCCAGTTTCTTCTTCTAATTCTTCTCTTAAACATGAACCTTCTCTTC 460

pCEN1-1 TGTCT----- 51
 pCEN1-4 TGTCT----- 68
 pCEN1-9 TGTCT----- 185
 B1.4E1.5 TGTCTGTAAGTTTAGTTGTCTTAATCCGTGAGTCTTAATTAGACATTGAAATTGTTGTG 400

pCEN1-5 ttagatTTTgatatctattggaactcttactagctcattcatccgtc 47
 pCEN1-8 tttgatatctattggaactcttactagctcattcatccgtc 41
 pCEN1-1 -----
 pCEN1-4 -----
 pCEN1-9 -----
 B1.4E1.5 GAAGCTGATGATCTTTAACTCTGTTTATTATGTGACTCACTTGTGCTTTTGTTAATTTT 340

pCEN1-5 ggCCTGTGAATTGTGTTTCTGAAGTTTCCCCAATGGCTGCTTTTGGAGATGACATTGTGA 107
 pCEN1-8 ggCCTGTGAATTGTGTTTCTGAAGTTTCCCCAATGGCTGCTTTTGGAGATGACATTGTGA 101
 pCEN1-1 --CCTGTGAATTGTGTTTCTGAAGTTTCCCCAATGGCTGCTTTTGGAGATGACATTGTGA 109
 pCEN1-4 --CCTGTGAATTGTGTTTCTGAAGTTTCCCCAATGGCTGCTTTTGGAGATGACATTGTGA 126
 pCEN1-9 --CCTGTGAATTGTGTTTCTGAAGTTTCCCCAATGGCTGCTTTTGGAGATGACATTGTGA 243
 B1.4E1.5 AGCCTGTGAATTGTGTTTCTGAAGTTTCCCCAATGGCTGCTTTTGGAGATGACATTGTGA 280

pCEN1-5 TTGTAGC-----GGCATATCGTACCGCCATTGCA 138
 pCEN1-8 TTGTAGC-----GGCATATCGTACCGCCATTGCA 132
 pCEN1-1 TTGTAGC-----GGCATATCGTACCGCCATTGCA 140
 pCEN1-4 TTGTAGC-----GGCATATCGTACCGCCATTGCA 157
 pCEN1-9 TTGTAGC-----GGCATATCGTACCGCCATTGCA 273
 B1.4E1.5 TTGTAGCGTAAGT... <= (65 bp) => ...CTTCAGGGCATATCGTACCGCCATTGCA 158

pCEN1-5 GCGAGACGTGGAGGTTTCAAAGACACTCTTCCTGATGATCTTCTTGCTTCTGTTCTTAAG 200
 pCEN1-8 GCGAGACGTGGAGGTTTCAAAGACACTCTTCCTGATGATCTTCTTGCTTCTGTTCTTAAG 194
 pCEN1-1 GCGAGACGTGGAGGTTTCAAAGACACTCTTCCTGATGATCTTCTTGCTTCTGTTCTTAAG 202
 pCEN1-4 GCGAGACGTGGAGGTTTCAAAGACACTCTTCCTGATGATCTTCTTGCTTCTGTTCTTAAG 219
 pCEN1-9 GCGAGA..... 281
 B1.4E1.5 GCGAGACGTGGAGGTTTCAAAGACACTCTTCCTGATGATCTTCTTGCTTCTGTTCTTAAG 160

pCEN1-5 AA-----GCTGTAGTGGAAAGAACATCTTTGGATCC 229
 pCEN1-8 AA-----GCTGTAGTGGAAAGAACATCTTTGGATCC 223
 pCEN1-1 AA-----GCTGTAGTGGAAAGAACATCTTTGGATCC 231
 pCEN1-4 AA-----GCTGTAGTGGAAAGAACATCTTTGGATCC 248
 B1.4E1.5 AAGTAAGT... <= (69 bp) => ...ATTAGGCTGTAGTGGAAAGAACATCTTTGGATCC 1

BamH I

determine the sequence of the isolated cDNAs. To this end, a phagemid representative of each of the two groups of cDNAs was selected for complete sequencing of its insert. The clone pCEN1-9 was chosen for being the longest of its group and, therefore, the most likely to correspond to a full-length transcript. For the second group, pCEN1-8 was preferred to pCEN1-5 as the restriction and hybridization patterns indicated it should harbour a single cDNA insert (see below).

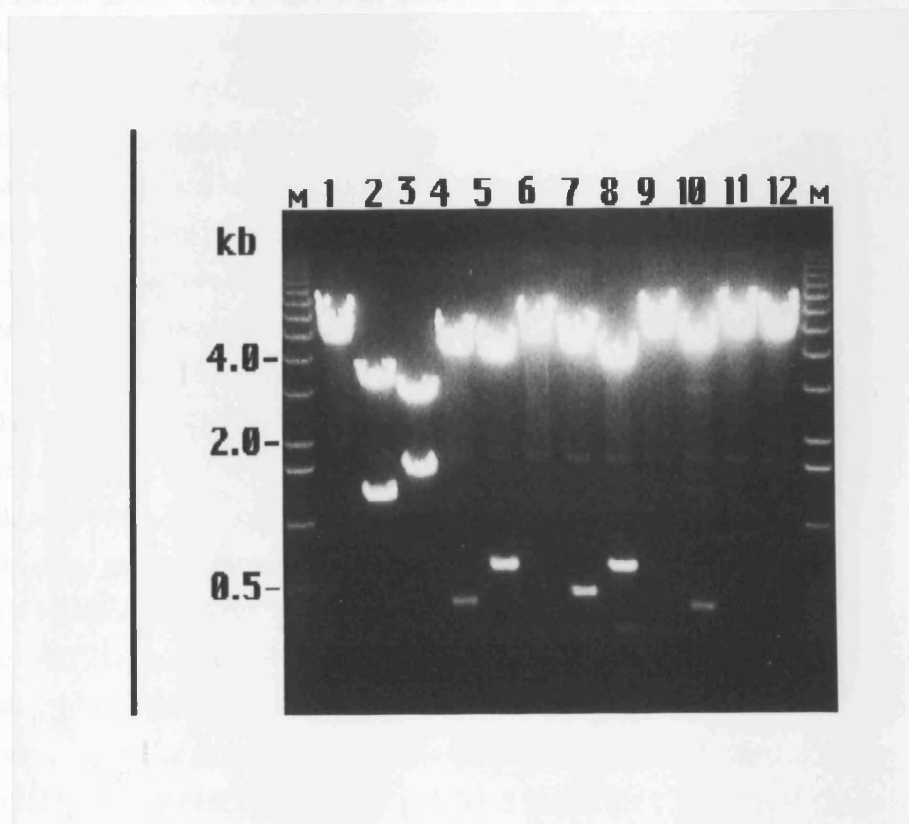
Simple restriction maps were constructed for both pCEN1-8 and pCEN1-9 in preparation for sequencing. Initially, pCEN1-9 DNA was digested with each of a battery of enzymes including *Acc* I, *Bam*H I, *Eco*R I, *Eco*R V, *Hinc* II, *Hind* III, *Kpn* I, *Pst* I, *Sma* I, *Sst* I, *Xho* I and *Xba* I to determine which would be most useful to generate subclones of the insert. After fractionation in an agarose gel it was possible to map most of the sites located in the insert taking into account its orientation relative to the vector (Fig. 5.2). The only exception concerned a *Pst* I site as there were two such sites in the insert separated by 0.6 kb. This *Pst* I site was, however, present in the previously determined partial sequences (Section 5.2.1) 204 bp upstream from the polyadenylation site in pCEN1-9. The restriction pattern obtained for pCEN1-8 with the enzymes *Bam*H I, *Hinc* II and *Pst* I suggested that its insert was essentially identical to that of pCEN1-9 (Fig. 5.2). It was therefore likely that the two cDNAs differed only at their 5' ends and their polyadenylation sites.

For sequencing the cDNA in pCEN1-9, subclones of it were generated by digesting the phagemid with each of *Bam*H I, *Hinc* II, *Pst* I and *Kpn* I. The resulting fragments were resolved by electrophoresis, individually isolated and purified from the agarose gel. For cloning, those fragments containing the vector sequence were self-ligated. Other fragments were ligated into the vector pBluescript KS (+) previously digested with the appropriate enzyme. The various ligation mixtures were transformed into *E. coli* as described in Section 2.2.2. Suitable clones were selected by preparing plasmids from transformants and verifying that these harboured the correct inserts by restriction enzyme analysis. Sequencing was then performed manually with the universal forward and/or reverse primers. For complete sequencing of the 0.6 kb *Pst* I fragment it was necessary to use two sequence-specific oligonucleotide primers. The primers were prTHIOP-F (GTTTGGCTATGAAGAAGGG) and prTHIOP-R (AGTACACATACTGAGATGC). For convenience and to facilitate subsequent work several other specific primers were designed and used both for manual and automated sequencing. The primers utilized included, prTHIO-1R (ACACGAGCGTTGGACAAGT-TATC), prTHIO-2F (AAGATATGGCTGCGGTGGAGTCTC), prTHIO-3R (CTCTTACTAG-CTCATTCATCCG) and prTHIO-4R (CAGTTGACATTGATTCCACTCC). The strategy followed for sequencing is shown in schematic form in Figure 5.3A. Sequencing of the cDNA in pCEN1-8 did not require subcloning. It was performed by a combination of manual and automated methods utilizing the phagemid DNA and several of the sequence-specific primers designed from the pCEN1-9 sequence, as diagrammed in Figure 5.3B.

Figure 5.2 Restriction enzyme maps of pCEN1-8 and pCEN1-9.

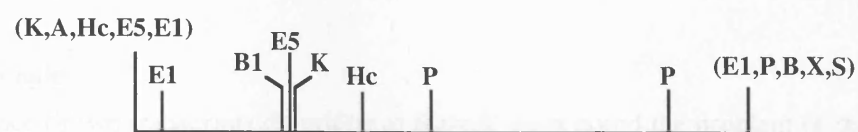
- A) Agarose gel electrophoresis of restriction enzyme digests of pCEN1-9 DNA. DNA was prepared as described in Section 2.4.1, and digested with various restriction enzymes. Lane M: 1 kb DNA marker ladder; Lane 1: *Acc* I; Lane 2: *Bam*H I; Lane 3: *Eco*R I; Lane 4: *Eco*R V; Lane 5: *Hinc* II; Lane 6: *Hind* III; Lane 7: *Kpn* I; Lane 8: *Pst* I; Lane 9: *Sma* I; Lane 10: *Sst* I; Lane 11: *Xba* I; Lane 12: *Xho* I.
- B) Restriction map of pCEN1-9.
- C) Restriction map of pCEN1-8.

A)

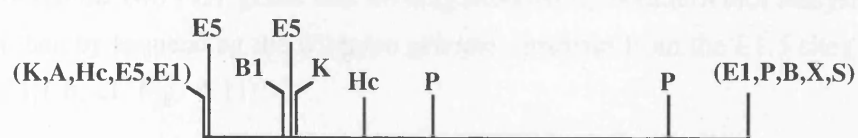


B) pCEN1-9

200 bp



C) pCEN1-8



The insert of pCEN1-8 was 1450 bp-long and that of pCEN1-9 1610 bp, their poly(dA) tails contributing, respectively, 19 and 16 residues to their total length. The two cDNAs only differed at their 5' and 3' ends as had been shown by the partial sequencing analysis (Section 5.2.1). Identity between the sequences started at positions 64 in pCEN1-8 and 196 in pCEN1-9. The polyadenylation site of pCEN1-9 was 21 bp downstream from that found in pCEN1-8. Translation of their sequences showed that each contained a single long Open Reading Frame (ORF) internal to and extending for most of the length of the cDNAs, as is reported in Section 7.2. The ORFs were found to be nearly identical, that of pCEN1-9 only differing from the ORF of pCEN1-8 in having an N-terminal extension of 43 codons (see Section 7.2). Structural characterization and comparative analysis showed that the ORFs were highly similar to 3-ketoacyl-CoA thiolases, as discussed in Chapter 7. The corresponding genes were designated *PKT1* (pCEN1-8) and *PKT2* (pCEN1-9).

The remaining cDNA clones were also completely or partially sequenced to determine if they could represent additional classes of transcripts. Sequencing was performed mostly on a single strand using several of the available primers (Fig. 5.3). The two shorter cDNAs in pCEN1-2 (1114 bp) and pCEN1-7 (954 bp) were found to correspond to 5' truncations of either the *PKT1* or *PKT2* transcripts (see Fig. 5.3E and F). The partial sequence data obtained on pCEN1-5 showed that the 1.5 kb fragment cDNA was identical to that in pCEN1-8. The 317 bp sequenced at the 3' end of pCEN1-10 were identical to the 3' end of pCEN1-8. The partial sequences of the 0.6 kb *EcoR* I fragment of pCEN1-5 and of the 1.2 kb *EcoR* I fragment of pCEN1-10, did not share any similarity with either pCEN1-8 and pCEN1-9, or to the genome in the T-DNA tagged region. They are likely to correspond to cDNAs fortuitously cloned together in the same vector during construction of the cDNA library.

5.3 Overlapping of the *PKT1* and *PKT2* transcription units

5.3.1 Rationale

The existence of two transcripts dissimilar at their 5' ends posed the problem of their origins. This was particularly true in view of the apparent orientation of the transcripts relative to the *gus* gene in AtEN101 (Section 5.2.1). Whereas *PKT2* was almost certain to originate from the T-DNA tagged region, the distinct 5' end of the *PKT1* transcript indicated that it might derive from elsewhere in the genome. A more likely alternative, given the absolute identity between the transcripts for most of their length, was the overlap of the *PKT1* and *PKT2* genes. The relation between the two *PKT* genes was investigated first by Southern blot analysis (Section 5.3.2), and then by sequencing the wildtype genome upstream from the E1.5 site (in the direction of E1.6, cf. Fig. 4.11).

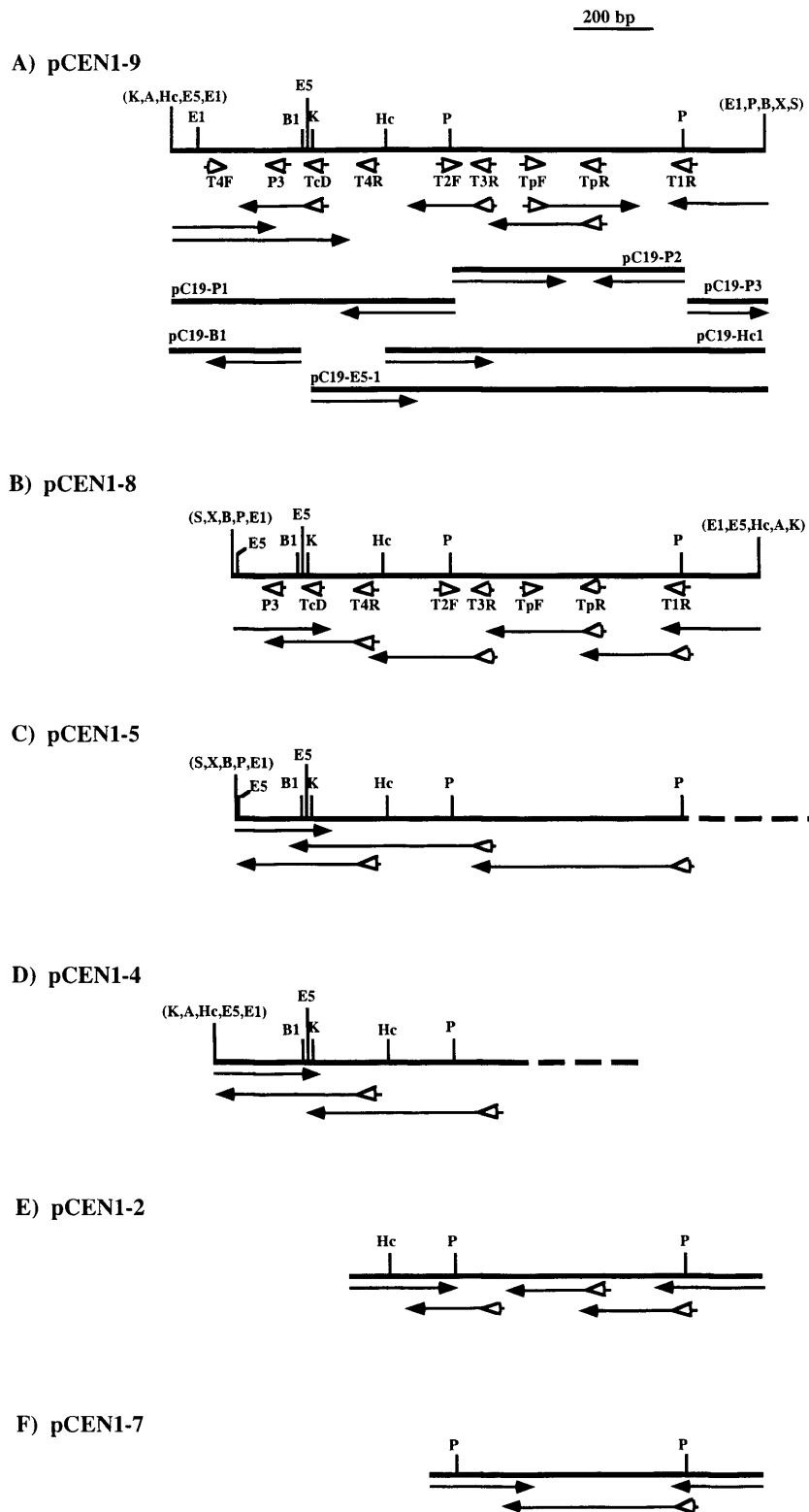


Figure 5.3 Sequencing strategies for the cDNA clones pCEN1-x.

Thick bars indicate the length of the inserts in the various clones and subclones. Arrows below the bars indicate the location, length and direction of the determined sequences. Discontinuous bars indicate non-sequenced segments of the inserts. Sequencing was either manual or automatic. Primer sequences are represented by open arrowheads. The primers utilized included prTHIO-cDNA (TcD), prTHIO-P-F (TpF), prTHIO-P-R (TpR), prTHIO-1R (T1R), prTHIO-2F (T2F), prTHIO-3R (T3R), and prTHIO-4R (T4R). Also shown are the positions of prEN1-P3 (P3) and prTHIO-4F (T4F).

5.3.2. Restriction map of the *PKT1* gene in the wildtype genome

If *PKT1* overlapped *PKT2*, its restriction pattern should match that of the wildtype genome in the tagged region. To establish the restriction pattern of the gene, samples of genomic DNA of wildtype *A. thaliana* cv. C24 and of AtEN101 were digested with *Bam*H I, *Eco*R I, *Eco*R V, *Hinc* II and *Hind* III. The products were resolved in an agarose gel (Fig. 5.4A) and a Southern blot prepared from it. Unfortunately the AtEN101 samples were not digested (Fig. 5.4A) undoubtedly because of the presence of some unknown contaminant. For hybridization, the cDNA in pCEN1-8 was isolated and purified following digestion with *Eco*R I and separation of the products by electrophoresis. A probe labelled with [α - 32 P]dCTP was obtained from the cDNA fragment by random oligonucleotide priming (Section 2.8.1). The probe was hybridized with the Southern blot at 65°C. After washing the blot according to standard procedures (Section 2.8.3), autoradiography was performed. The autoradiograph obtained is shown in Figure 5.4B, and the sizes of the wildtype genomic fragments hybridizing with the probe from the *PKT1* cDNA are listed in Table 5.1.

The hybridization results were indicative of an overlap between *PKT1* and *PKT2*. In particular, the single band of hybridizing *Eco*R I fragments and the only band of strongly hybridizing *Bam*H I fragments matched the patterns expected from the T-DNA tagged region as their sizes were identical to those of B1.3B1.4 and E1.3E1.4 (cf. Fig. 4.11). The apparent lack of hybridization to other fragments should be related to the short length (197 bp) of the pCEN1-8 cDNA extending upstream from its (putative) B1.4 site.

<i>Bam</i> H I	<i>Eco</i> R I	<i>Eco</i> R V	<i>Hinc</i> II	<i>Hind</i> III
> 12 kb*	4.6 kb	4.9 kb	2.3 kb	> 12 kb
3.4 kb	-	1.1 kb*	1.2 kb*	~ 12 kb
-	-	-	-	0.8 kb*

Table 5.1 Genomic fragments hybridizing to the *Eco*R I cDNA insert of pCEN1-8.

Listed are the sizes of the fragments detected in the Southern blot shown in Figure 5.4. Bands characterized by a weak hybridization signal are marked with an asterisk.

The same may be argued to explain the poorer hybridization of the 1.1 kb *Eco*R V and the 1.2 kb *Hinc* II fragments. There was also good agreement between the size of the hybridizing 1.1 kb *Eco*R V fragment and the known distance (1063 bp) between the E5.1 site (Fig. 4.11) and the *Eco*R V site located 13 bp from the B1.4 site in one of the B1.3B1.4 partial sequences (Section 4.3.4). The total length of overlap between the two sequences was 196 bp dispersed by three exons.

5.3.3 Construction of a restriction enzyme map for pENB8

Although strongly indicative of overlap between the two genes, the Southern blot data results could not offer final proof of it as they might have arisen from a high conservation of structure between two distinct genomic fragments harbouring *PKT1* and *PKT2*. Sequencing of the wildtype genome upstream from the E1.5 site should provide conclusive evidence with regard to this question. As a preparatory step for the sequencing, a restriction map was constructed for the 5.4 kb *Bam*H I fragment B1.4B1.5 subcloned in plasmid pENB8 (cf. Section 4.3.3).

In a preliminary experiment a number of restriction enzymes were tested for their ability to digest the fragment B1.4B1.5 of L-EN1-4. To this end pENB8 was digested with *Bam*H I. A second round of restriction enzyme digestions was carried out with each of *Apa* I, *Bcl* I, *Bgl* II, *Cla* I, *Eco*R V, *Hinc* II, *Hind* III, *Hpa* I, *Kpn* I, *Pst* I, *Sal* I, *Sca* I, *Sma* I, *Sst* I, *Sst* II, *Xba* I and *Xho* I. Of these, only *Eco*R V, *Hinc* II, *Hpa* I and *Xba* I were found to have sites in the segment (data not shown). To map sites close to the region of T-DNA insertion several single and double digests of pENB8 were prepared and separated in an agarose gel. The sizes of the generated fragments were estimated by comparison with those of a marker (Fig. 5.5A). The mapping took in consideration the already known partial sequences of the plasmid which had established the location of one of the *Eco*R V sites and the orientation of the insert relative to the flanking polylinker sites on the vector (Section 4.3.3). The resulting restriction map is shown in Figure 5.5B.

5.3.4 Sequencing the E5.1E5.2 genomic fragment

The fragment E5.1E5.2 was isolated from pENB8 by the usual procedures (Section 2.4.6.1) and ligated into *Eco*R V-linearized pBluescript KS(+). DNA of plasmid pE5.1E5.2-2 from a selected transformant containing a single copy of the E5.1E5.2 segment was prepared and sequenced in both directions with the universal primers. The sequence was found to be 298 bp-long making the B1.4E5.2 segment of genome 1342 bp in length. The E5.1E5.2 sequence was compared to those of the pCEN1-5 and pCEN1-8. The 5' ends of the cDNAs in these two clones, that had failed to match the sequence of the B1.4E1.5 fragment (Section 5.2.1), aligned with that of E5.1E5.2 (Fig. 5.6A).

The results demonstrated that the two transcriptional units, *PKT1* and *PKT2*, overlapped each other and shared most of their genic structure. The first exon represented in pCEN1-5 and pCEN1-8 was thus separated from the second exon by a long intron spanning 966 bp. Located within this intron was the first exon of the pCEN1-9-type cDNAs, as well as the T-DNA insertion point in AtEN101. A diagram of the deduced structure is shown in Figure 5.6B.

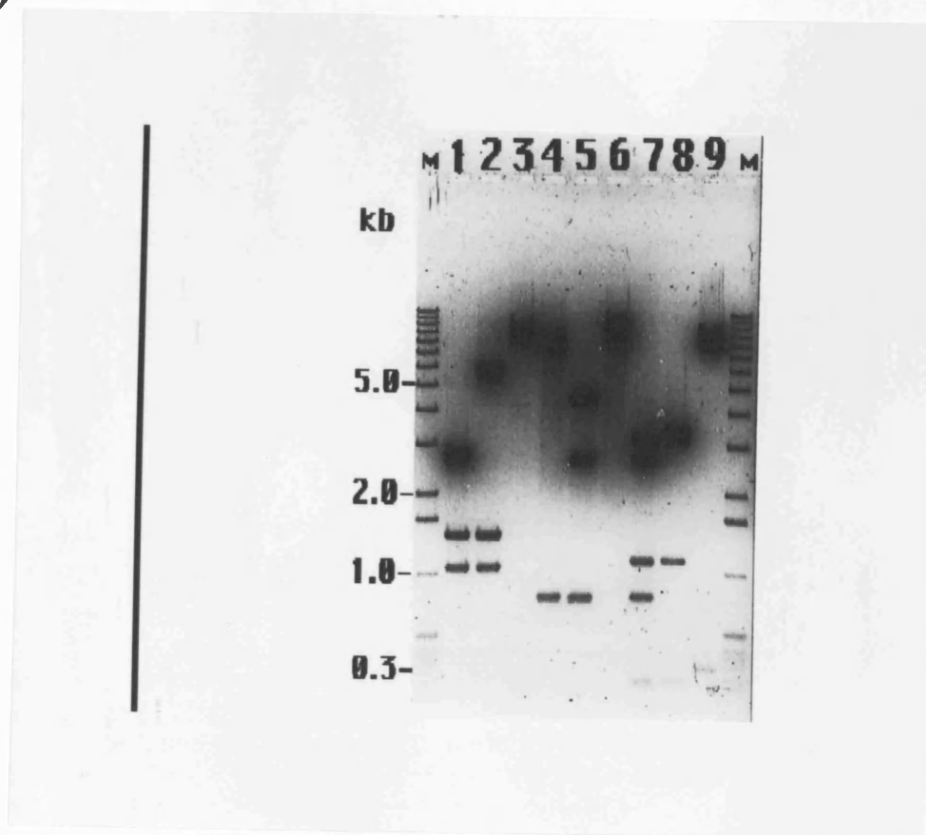
5.3.5 Searching for additional exons of *PKT2*

In addition to the possible overlap between two genes, the two transcriptional units *PKT1* and *PKT2* represented in the cDNAs could derive from differential splicing of a single gene. Such

Figure 5.5 Mapping restriction enzyme sites in pENB8.

- A)** Aliquots of the plasmid digested with various enzymes were resolved in the agarose gel shown. Lane M: 1 kb DNA ladder; Lane 1: *Bam*H I + *Xba* I; Lane 2: *Xba* I; Lane 3: *Kpn* I; Lane 4: *Kpn* I + *Hpa* I; Lane 5: *Bam*H I + *Hpa* I; Lane 6: *Hpa* I; Lane 7: *Bam*H I + *Hinc* II; Lane 8: *Hinc* II; Lane 9: *Eco*R V. A negative of the picture is shown as it is easier to visualize the smaller fragments in it.
- B)** Restriction map of the insert in pENB8. The nomenclature of the sites follows that of the restriction map in Figure 4.11. B1: *Bam*H I; E1: *Eco*R I; E5: *Eco*R V; Hc: *Hinc* II; X1: *Xba* I. Dimmed writing indicates sites previously mapped (Chapter 4). In bold are sites mapped from analysis of the gel in A).

A)



B)

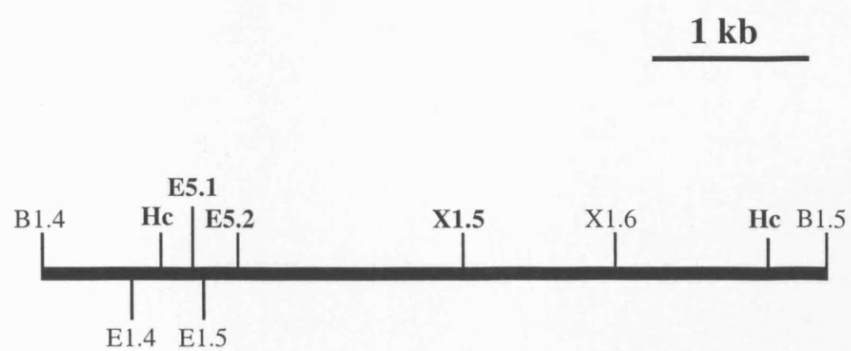


Figure 5.6 Overlap of the *PKT1* and *PKT2* genes in the tagged region.

- A)** Alignment of the 5' end of the cDNA in pCEN1-8 with the genomic sequence downstream from the *EcoR* V site E5.2, and with the 5' end of the cDNA in pCEN1-9. The E5.2 site is highlighted by double underlining. The match to the 5' splice site dinucleotide GT is also marked by italics and underlining. The genomic sequence has the direction of native transcripts (*i.e.* the reverse of that of the reporter gene and of the genomic sequence in Appendix 1). Its numbering increases in the direction of the downstream *BamH* I site B1.4.
- B)** Diagram showing the relative positions of the 5' exons of the transcripts corresponding to the pCEN1-8 and pCEN1-9 type cDNAs. The site of T-DNA insertion in AtEN101 is indicated.

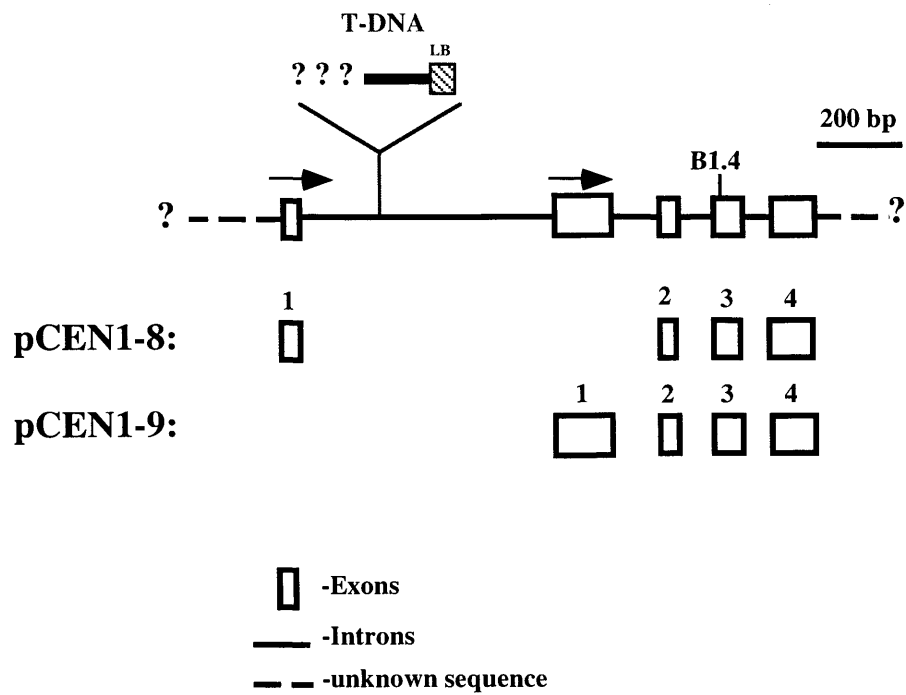
A)

```

pCEN1-9 (180)                               ...TCTCTTCTGTCTCCTGTGAATTGT...
                                         : : |||||
pCEN1-8 (1) TTTGATATCTATTGGAACTCTTACTAGCTCATTCATCCGTCGGCCTGTGAATTGT...
           |||||
E5.1E5.2 (1) GATATCTATTGGAACTCTTACTAGCTCATTCATCCGTCGGGTAACTTTCTTG...

```

B)



scenario would require that both types of primary transcripts shared the same (or nearly) 5' end. This would also require that the putative exon 1 of *PKT2* was preceded by at least one other exon. The possible existence of 5' and 3' splice sites in the region separating the known 5' ends of the two transcript types was investigated by searching for matches to the consensus, as described in Section 4.2.6.2. No matches were found in the region.

5.4 Approaching the 5' ends of the *PKT1* and *PKT2* transcripts by 5' RACE

5.4.1 Rationale

The divergent orientation of the thiolase genes *PKT1* and *PKT2* relative to the *gus* reporter (Section 5.2.1) and the failure to detect any other transcripts in the region (Section 4.4) raised questions with regard to the nature of the fusion transcription unit in AtEN101. As suggested in Section 4.2.5, the lack of detection of other transcripts mapping to the region might derive from their low abundance which could have prevented the identification of corresponding cDNAs during the library screening. A further plausible explanation would have been the presence of a cryptic promoter in the mutant line. The relative proximity of the 5' ends of the pCEN1-9-type cDNAs to the site of T-DNA insertion suggested that any cryptic promoter that might be responsible for the expression of *gus* in AtEN101 could be related to the *cis*-elements controlling the expression of *PKT2*. An initial investigation of this possibility could be achieved by determining and comparing the locations of the promoters both of the fusion transcripts in AtEN101 and of the *PKT2* transcripts by mapping their 5' ends. This could also clarify if the *PKT1* and *PKT2* transcripts were splicing variants of a single gene.

With regard to the *PKT2* gene, as well as *PKT1*, the actual 5' ends of their transcription units were unknown although pCEN1-8 and pCEN1-9 contained complete ORFs (Section 7.2). The genomic regions immediately upstream from the 5' end of the *PKT2* transcripts represented in the cDNAs did not show obvious promoter characteristics. The only potential regulatory signal found was a CCAAT box 117 bp upstream from the 5' end of the cDNA in pCEN1-9 (cf. Fig. 8.9).

The approach adopted to map the promoter regions of the fusion and *PKT1/PKT2* transcripts consisted of cloning their upstream terminal segments following 5' Rapid Amplification of cDNA Ends (5' RACE). The concept of 5' RACE refers to the reverse transcription into the unknown 5' region of a target transcript and subsequent modification of the 3' end of the new cDNA by the addition of a short extension of known sequence (Belyavsky *et al.*, 1989; Frohman *et al.*, 1988). The 5' end of the modified cDNA is then amplified by PCR with oligomeric primers specific for the known sequence of the cDNA and for the 5'-terminal extension. The sequence of the amplified extension can then be determined directly or following cloning. The 5' RACE of the *PKT1* and *PKT2* transcripts is the subject of the

subsequent Sections. Although performed in parallel, the application of 5' RACE to the fusion transcripts is reported in Chapter 6.

5.4.2 Applying the SLIC method to the *PKT1* and *PKT2* transcripts

5.4.2.1 Isolation of mRNA

Initially, a modification of the 5' RACE technique based on single-stranded ligation of an anchor sequence to single-stranded cDNA, SLIC (Edwards *et al.*, 1991), was adopted (Section 2.7.4.1), utilizing a 5'-AmpliFINDER™ kit (Clontech, California). Total RNA was isolated from siliques of wildtype plants grown in a greenhouse at 25°C under a day/night regimen (Section 2.6.2). Given that the cDNA clones were isolated from a library constructed with RNA from *A. thaliana* cv. Columbia, plants of the same variety were utilized to prepare the wildtype RNA. The quality of the RNA was evaluated by electrophoresis in a denaturing agarose gel and found to be intact (data not shown). Isolation of mRNAs from the total RNAs was performed with poly(dT)-coated-magnetic beads supplied in a "Dynabeads polyA⁺ RNA direct isolation" kit (Dyna, Norway), as described in Section 2.6.4. About 2 µg of mRNA recovered from a sample of total RNA were utilized for reverse transcription with sequence-specific primers (Section 2.7.4.1). 5' RACE was performed in parallel on both the wildtype mRNA and on a supplied control RNA sample according to the manufacturer's instructions provided with the kit. The control RNA, of placental origin, was used to synthesize a cDNA for the human transferritin receptor with a primer also supplied with the kit.

5.4.2.2 SLIC of *PKT1* and *PKT2* transcripts

The primer utilized for reverse transcription of the *PKT1*/*PKT2* transcripts, prEN1-P3 (GCAAGAAGATCATCAGGAAGAGTGTC), was designed from the common portion of their sequences. It mapped at 175 bp and 297 bp downstream from the 5' ends of the cDNAs in pCEN1-8 and pCEN1-9, respectively. The newly synthesized single stranded cDNA was purified according to the recommended protocol and then ligated to the single-stranded oligomeric AmpliFINDER anchor supplied with the kit. For single-stranded anchor ligation T4 RNA ligase was used. The supplied AmpliFINDER anchor is blocked at its 3' end to prevent concatamer formation. For PCR amplification of the modified cDNA, a supplied anchor-specific AmpliFINDER primer and a sequence-specific primer were used. The latter was prEN1-P1 (AACCTCCACGTCTCGCTTTGC), nested 7 bp to prEN1-P3. After 36 cycles of PCR amplification 15 µl samples were separated on an agarose gel. Lack of any visible products following electrophoresis led to a re-amplification of an aliquot of the first PCR reaction. For this step a third primer was designed, prEN1-P2 (CGCTACAATCACAAT-

GTCATCTCC), nested 96 bp to prEN1-P1. No products of amplification could be seen when samples were run on an agarose gel.

The absence of PCR products could be due to a variety of factors, among them the reaction conditions. The PCR had been conducted as recommended by the kit suppliers and had been found to be effective when tested for the amplification of an internal fragment of pCEN1-9 with each of the two primers prEN1-P1 and prEN1-P2 (for both tests the conjugate primer had been prTHIO-4F mapping to positions 52-79 in the cDNA sequence). Nevertheless, PCR amplification of RACE cDNA was repeated under various other combinations of annealing temperatures and extension times, but always unsuccessfully.

5.4.2.3 SLIC of the control human transferritin RNA

To determine the causes of the failure in obtaining RACE products, samples of the synthesized human control cDNA taken prior to and after its purification, and also following cDNA ligation to the AmpliFINDER anchor, were amplified by PCR. The first two samples were amplified with two supplied control-specific primers (TFR2/TFRQC) which should generate a 338 bp product. One of these primers was replaced with the anchor primer to test the effectiveness of the ligation reaction. For this sample the size of the expected amplified fragment was 480 bp. Agarose gel electrophoresis of aliquots of the various reactions showed strong bands of the predicted size for the TFR2/TFRQC reactions (Fig. 5.7). However, the ligation reaction had been very inefficient as only a trace of the expected size products could be seen (lane 3; Fig. 5.7).

5.4.3 Development of a new 5' RACE system (LOPRACE)

The very considerable difficulties found with the SLIC method prompted the search for an alternative approach to clone the 5' ends of the *PKT1/PKT2* transcripts. A long established property of reverse transcription reactions is the folding back of the growing cDNA chain when it reaches the 5' end of the template transcript (Kimmel & Berger, 1987). This phenomenon was widely utilized in the earlier days of construction of cDNA libraries to prime second strand synthesis. It was thought that this property could be exploited to perform 5' RACE with a single primer or a nested primer oriented in the same direction. This loop-primed RACE (LOPRACE) procedure, schematized in Figure 5.8, consists of, briefly:

1. Reverse transcription with a sequence-specific primer (Fig. 5.8A).
2. The 3' looping end of the cDNA strand in the RNA-DNA hybrid is used to prime second strand synthesis generating a long cDNA hairpin (Fig. 5.8B).

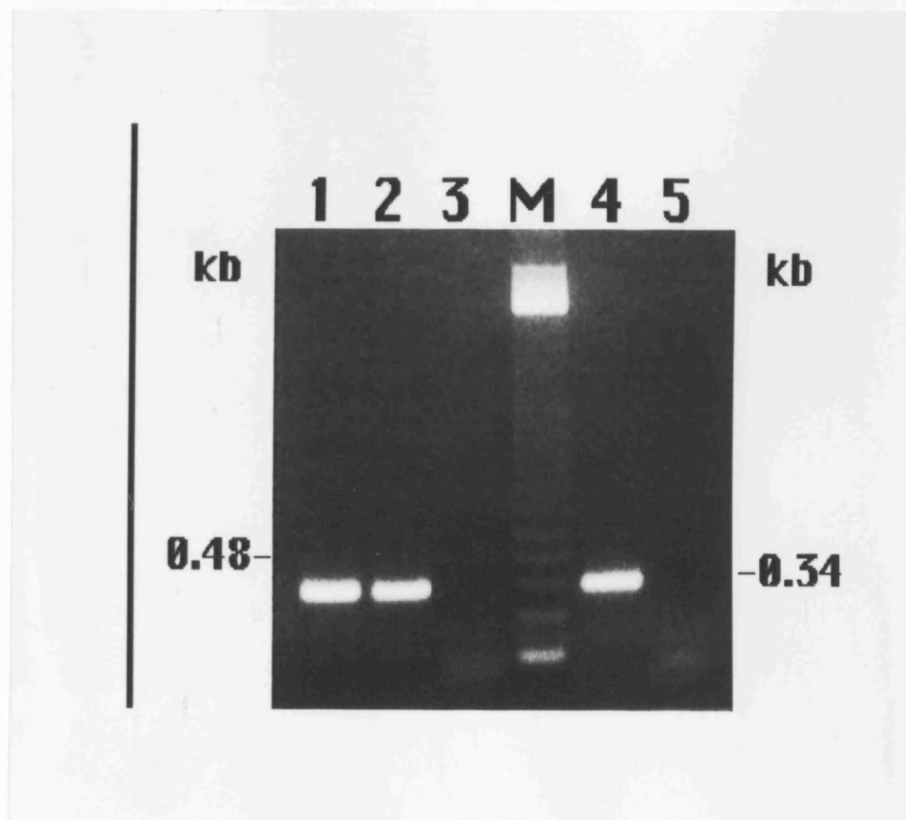
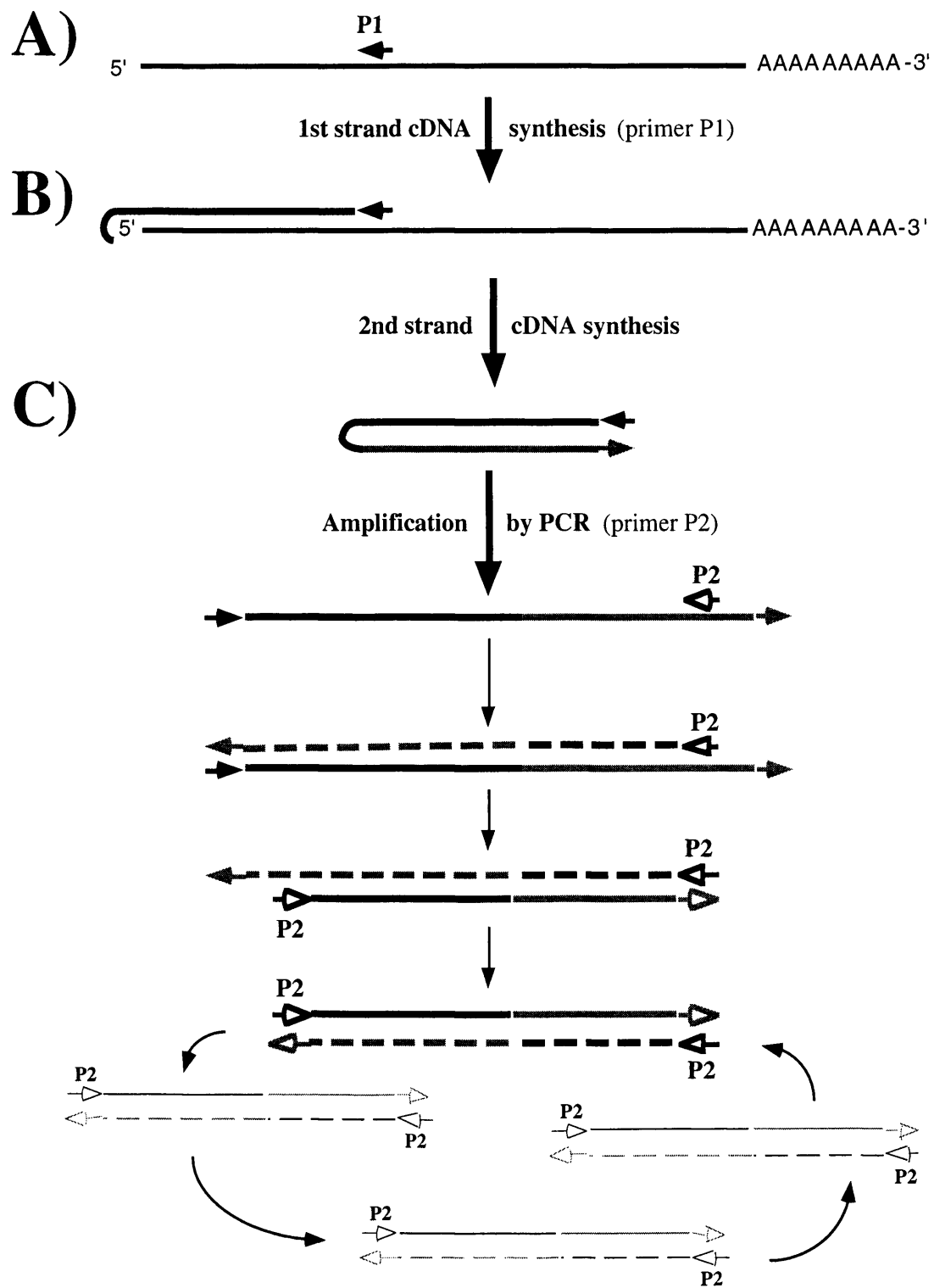


Figure 5.7 Control of SLIC procedure.

The SLIC procedure was applied to a control human RNA supplied with the 5'-AmpliFINDER™ kit (Clontech, California). Samples were taken at different steps of the procedure and submitted to PCR amplification with sequence-specific primers. The amplification products were separated in an agarose gel. Lane 1: Sample from cDNA synthesis reaction prior to purification (primers TFR2 and TFRQC); Lane 2: cDNA following purification, non-ligated (primers TFR2 and TFRQC); Lane 3: cDNA ligated to anchor (primers TFR2 and AmpliFINDER); Lane 4: cDNA ligated to anchor (primers TFR2 and TFRQC); Lane 6: Negative PCR control including all three primers; Lane M: 1 kb DNA marker. Amplifications were performed as detailed in Section 2.7.4.1. Cycling was 25 x (94°C 45 secs, 60°C 45 secs, 72°C 2 min) for the TFR2/TFRQC pair, or 35 x (94°C 45 secs, 60°C 45 secs, 72°C 2 min) with the TFR2/AmpliFINDER primer pair.

Figure 5.8 Diagram flow of LOPRACE procedure.

- A) First strand cDNA synthesis by reverse transcription with a sequence-specific primer (P1).
 - B) Second strand cDNA synthesis by extension of the loop formed at the 3' end of the first strand cDNA. An hairpin structure is formed.
 - C) PCR amplification of the cDNA using a single sequence-specific primer (P2) nested to P1. The expected products at the end of the extension periods of successive PCR cycles are shown.
- Thick black bars indicate first strand cDNA sequences; Grey bars indicate second-strand (complementary) cDNA sequence. Continuous and discontinuous bars are used to distinguish the two annealed strands of DNA at the end of PCR extension reactions.



3. The hairpined cDNA is submitted to amplification by PCR with the same primer or a nested sequence-specific primer. The first cycle of the PCR will melt the hairpin and generate a complementary sequence of it. This complementary sequence is essentially a copy of the original. For this reason a single primer is required for the PCR reaction (Fig. 5.8C).

To explore the application of LOPRACE to the *PKT1/PKT2* transcripts, cDNAs were prepared from 1.0 µg of total RNA of *A. thaliana* cv. Columbia by reverse transcription with the primer prEN1-P3, as described in Section 2.7.4.2. A negative control lacking reverse transcriptase was processed in parallel. After adjusting the reaction buffer, second strand synthesis was performed with the Klenow fragment of the *E. coli* DNA polymerase. For amplification of the cDNAs the primer prEN1-P2 was utilized. Two rounds of amplification were necessary to visualize the products, the first lasting for 32 cycles and the second 28 cycles. Aliquots taken at the end of the second PCR cycling were resolved in an agarose gel and two bands of fragments were observed (Fig. 5.9). The larger fragments were estimated to be about 0.30 kb whereas the smaller were just below 0.20 kb.

For cloning, the secondary amplification reaction was repeated three times and the products pooled and concentrated. After electrophoretic resolution, the fragments were isolated from the agarose gel, purified and ligated to pCRTMII vector. Some *E. coli* transformants, given the designation pTM-x, were subjected to colony-PCR to estimate the sizes of the inserts in their plasmids (Fig. 5.10). Several of the plasmids were manually sequenced including pTM-2, -3, -4, and -7. Whereas the pTM-3 insert consisted of only pCRTMII vector and primer sequences, the remainder included sequences identical to different segments of the 5' end of the pCEN1-9 cDNA starting at the prEN1-P2 segment. Unlike the expected structure for these inserts they were found to terminate at an inverted prEN1-P2 sequence. Only in one of the clones, pTM-7, the sequence extended beyond the known 5' limit of the pCEN1-9 cDNA. This extension was 23 residues-long (Fig. 5.13) and closely matched the known genomic sequence in the region except for a 1 bp deletion (not shown). The total insert length for the pTM-7 clone was 297 bp and therefore seemed to correspond to the class of the largest fragments obtained by LOPRACE. A chromatogram obtained by automated sequencing of pTM-7 is shown in Appendix 2.

5.4.4 RACE by homopolymeric tailing

While developing LOPRACE, a modified version of the original 5' RACE technique (Frohman *et al.*, 1988) was employed for cloning of the *PKT1* and *PKT2* transcripts 5' ends (and also those of the fusion transcripts; see Section 6.2.3). For the application of this method, a "5' RACE System for Rapid Amplification of cDNA Ends" kit from Gibco/BRL was used. The procedures followed were those recommended by the manufacturer and are described in Section 2.7.4.3.

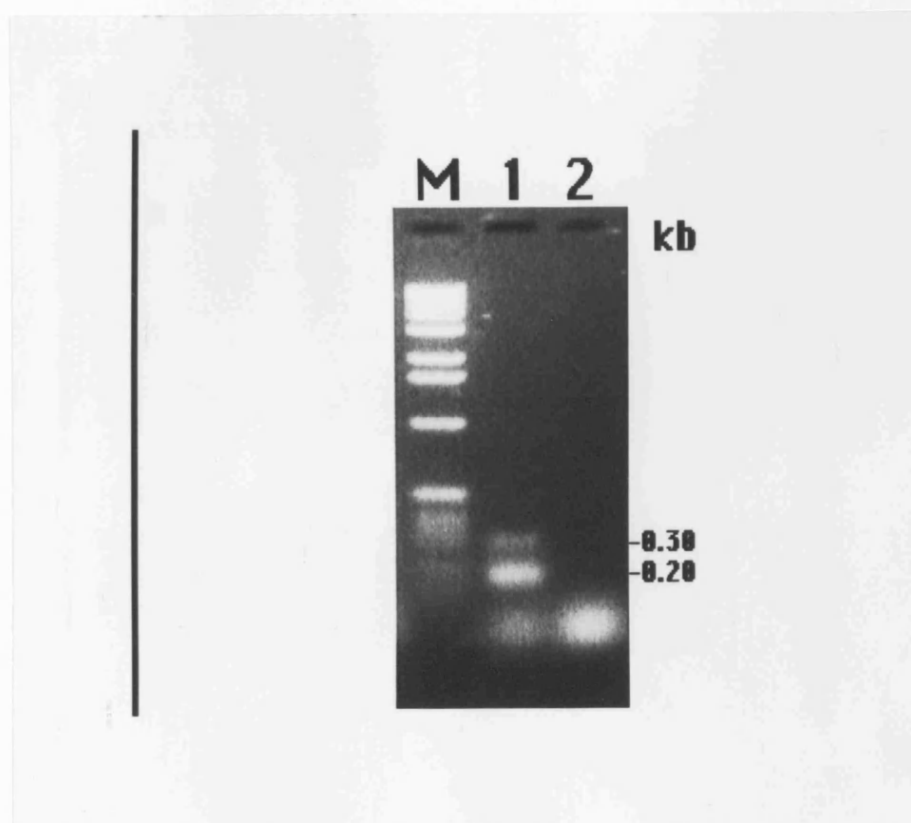


Figure 5.9 LOPRACE of *PKT1/PKT2* transcripts.

Visualization of products obtained after the second round of amplification with primer prEN1-P2. First-strand synthesis of cDNAs was carried out by reverse transcription with primer prEN1-P3 (Section 2.7.4.2). Lane M: 1 kb DNA marker; Lane 1: PCR of the cDNA sample; Lane 2: negative PCR control. For the negative control a mock cDNA synthesis reaction was conducted in parallel with the test reaction but omitting the addition of reverse transcriptase. An aliquot was then submitted to the same amplification protocol as the experimental sample. Amplification was 32 x (94°C 45 secs, 60°C 30 secs, 72°C 90 secs) followed by an additional round of 28 cycles with identical conditions.

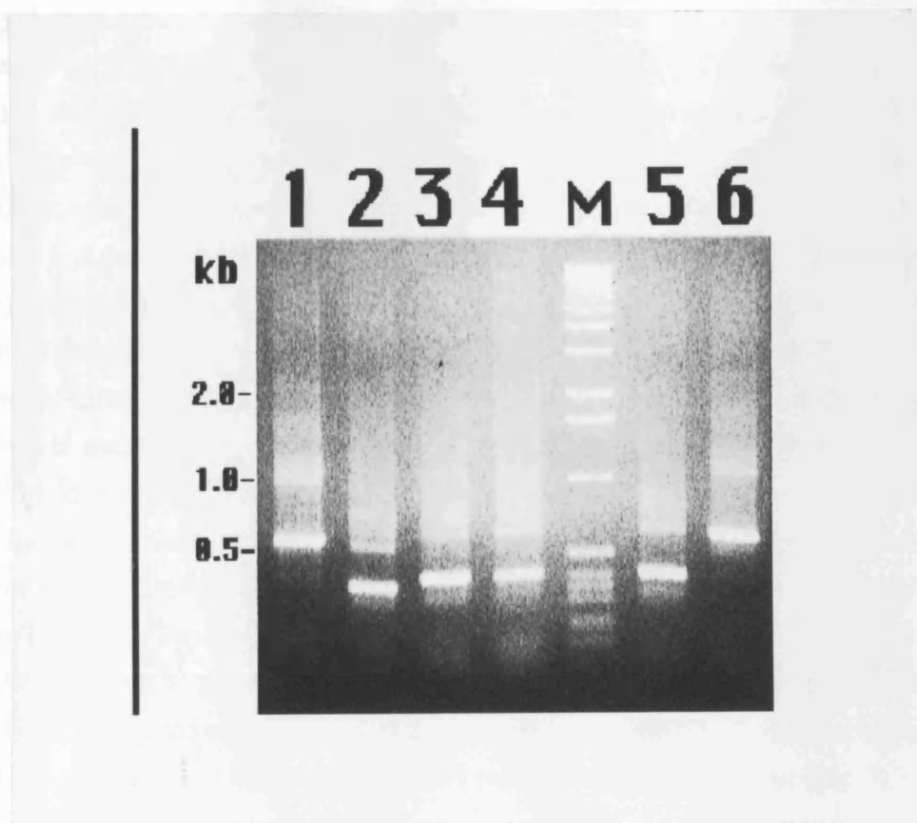


Figure 5.10 Colony-PCR of LOPRACE pTM clones for *PKT1/PKT2* transcripts.

Shown is the agarose gel resolution of the amplified inserts of several clones. Amplification was with the universal forward and reverse primers. Lane 1: pTM2; Lane 2: pTM3; Lane 3: pTM4; Lane 4: pTM5; Lane 5: pTM6; Lane 6: pTM7; Lane M: 1 kb DNA ladder (Gibco/BRL). The expected contribution of the vector sequences to the amplified fragments was 242 bp.

Total RNA of wildtype *A. thaliana* cv. Columbia was isolated from 10 day-old seedlings grown in 1/2MS₁₀ as described in Section 2.6.2. Its intactness was verified by agarose gel electrophoresis (not shown). The RNA was then used for first strand cDNA synthesis with prTHIO-5R (GCACCAATACCAATGTCGTAA). The primer was designed to overlap two exons of the *PKT* genes as deduced from a partial sequence of the B1.3B1.4 fragment in plasmid pENB3 (Section 4.3.4). Following its 3'-homopolymeric tailing with dC, the cDNA sample was subjected to PCR with the supplied anchor primer and with prEN1-P3 for 32 cycles. A second round of amplification lasting 20 cycles under the same conditions was necessary for visualization of products by agarose gel electrophoresis (Fig. 5.11A). Several bands of ethidium bromide-stained DNA could be seen (Fig 5.11A; lane 2). The strongest corresponded to the largest fragments of about 0.4-0.5 kb. Amplification of the cDNA with the anchor primer and the cDNA synthesis primer (prTHIO-5R) was also performed. The resulting products could be seen as smeared bands on the gel (Fig 5.11; lane 3). A negative control of cDNA synthesis processed in parallel ensured that the bands were not PCR artifacts (lane 1). A positive control of cDNA synthesis was also conducted by amplifying the cDNA with prTHIO-5R and prTHIO-4F. The two primers should amplify a 510 residues-long segment at the 5' end of the *PKT2* cDNAs. Fragments of this estimated size were observed as expected (Fig. 5.11A; lane 4). For cloning, the second round of amplification with prEN1-P3 and Anchor primers was repeated in large-scale and the pooled products separated in an agarose gel (Fig. 5.11B). The products were more finely resolved and it was observed that the 0.4 kb fragments (Fig 5.11A) mostly included fragments of 0.42 kb and 0.48 kb (Fig. 5.11B). Other less abundant products of smaller size were seen including fragments of about 0.20 kb and 0.25 kb. The DNA fragments of about 0.4 kb and 0.2-0.3 kb were separately extracted from the gel with a "High Pure PCR Product Purification" kit (Boehringer Mannheim). After purification the DNAs were ligated into pCRTMII vector and the products transformed into *E. coli*. Several transformants, designated pTB-x, were isolated. The sizes of the inserts in their plasmids were determined by restriction enzyme analysis with *Bst*X I for those derived from the 0.4 kb PCR products (Fig. 5.12; top). For the clones obtained with the smaller-size fragments colony-PCR was utilized to screen the transformants (Fig. 5.12; bottom).

Several of the plasmids corresponding to the longer (0.4 kb) PCR products, including pTB-3, -5, -6, -7, and -11, were chosen for sequencing. For the smaller PCR products, sequencing was performed on plasmids pTB-21, -22, -23, -24, -25, -27 and -28. Both manual and automated methods were used for sequencing together with the universal forward and reverse primers. The sequences obtained from the pTB plasmids with longer inserts corresponded to the transcripts of *PKT2* whereas *PKT1* seemed to be only associated with shorter PCR products. Several of the plasmids extended the known sequences for the two transcripts at their 5' ends. For *PKT1* only two RACE clones, pTB-21 and pTB-24 extended the known sequence at the 5' end of the transcripts (as in pCEN1-5), by 14 residues (Fig. 5.13).

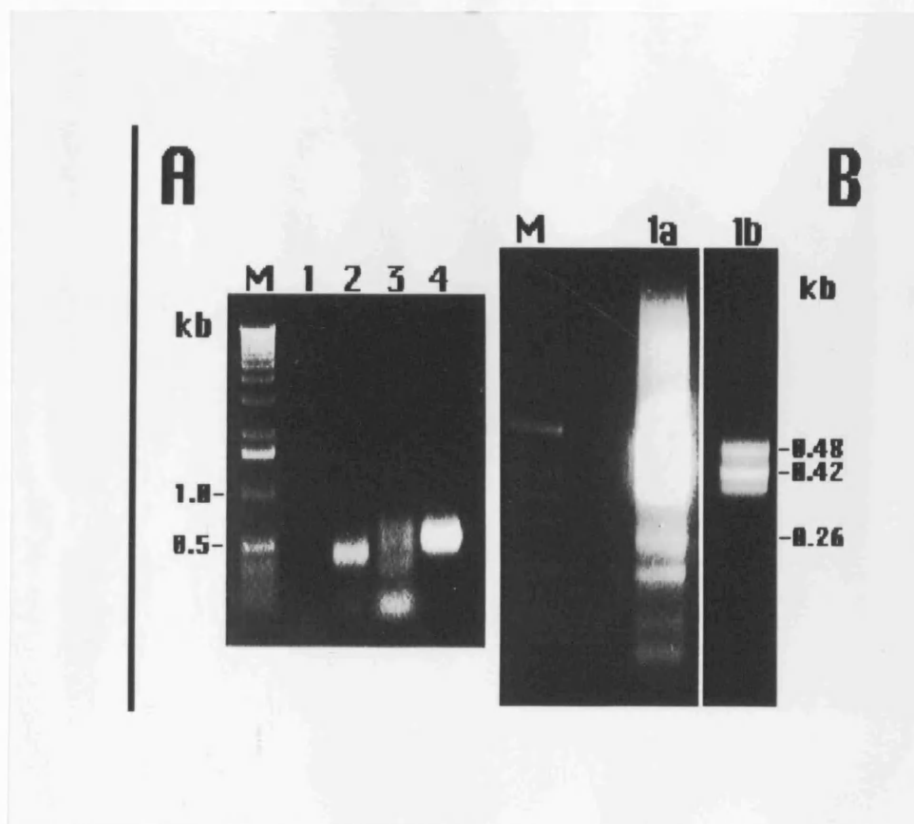


Figure 5.11 5' RACE of *PKT1/PKT2* transcripts (Homopolymeric tailing method).

- A)** Agarose gel of 5' RACE products of *PKT* transcripts. Three samples of dC-tailed first-strand cDNA were amplified with distinct sets of primers in two rounds of cycles. At the end of the second round of amplification 20 μ l aliquots were resolved in the gel shown. PCR conditions were as described in Section 2.7.4.3. Lane M: 1 kb DNA ladder marker. Lane 1: negative control of 5' RACE with prEN1-P3/Anchor primer. Reverse transcriptase was omitted from the cDNA synthesis reaction. Lane 2: cDNA amplification with prEN1-P3/Anchor primer. Lane 3: cDNA amplification with prTHIO-5R/Anchor primer. Lane 4: positive control of cDNA synthesis with primers prTHIO-5R and prTHIO-4F. The positive control is expected to amplify a 510 bp segment of a *PKT2*-derived cDNA.
- B)** Resolution of the products of the large-scale 5' RACE. Second round amplifications were performed with prEN1-P3/Anchor primer pair, pooled and separated in the agarose gel shown (Lane 1a). To permit a clear visualization of the fragments of about 0.4 kb, an identical but underexposed image of the same lane is shown (Lane 1b). The estimated sizes of the most abundant fragments are indicated.

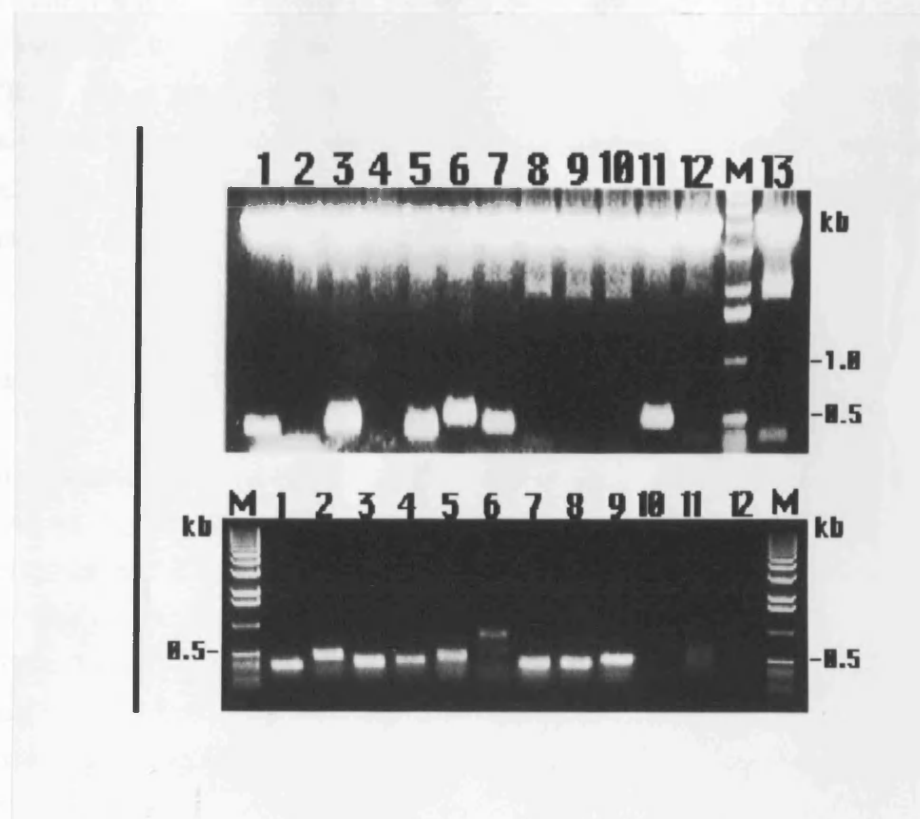


Figure 5.12 Sizing the inserts in the pTB-x plasmids.

Top: Agarose gel of pTB-x clones digested with *Bst*X I. Lane M: 1 kb DNA ladder marker. Lanes 1-13: pTB-1 to -13.

Bottom: Colony-PCR of pTB-x clones. PCR was with the universal forward and reverse primers. Lanes 1-12: clones pTB-20 to -32.

The 5' termini of the other four RACE clones also mapped to the first exon in pCEN1-5. The RACE strategy was also successful in extending the known 5' termini of *PKT2* transcripts in four of the clones. The extensions were of 32 residues for pTB-7, 49 residues for clones pTB-3 and pTB-11, and of 105 residues for clone pTB-6 (Fig. 5.13). Comparison with the genomic sequence in the region showed that the additional segments were continuous with the first exon of the cDNA in pCEN1-9. The resulting composite sequences for *PKT1* and *PKT2* were designated mP18 and mP19, respectively (Fig. 5.13).

5.5 Search for ESTs of the *PKT1* and *PKT2* genes

Current sequencing efforts for the molecular study of *Arabidopsis* have generated an increasing number of partial sequences of cDNAs. The availability of these sequences, commonly known as ESTs, at publicly accessible databases enabled their comparison with the pCEN1-8 and pCEN1-9 cDNAs. The search for similar sequences to the two cDNAs had two objectives including, in the context of the characterization of the *PKT1* and *PKT2* genes, extending the known sequence at their 5' ends. Determining if variants of the *PKT1* and *PKT2* transcripts or homologous genes existed constituted the additional objective, as is reported in Section 10.3.1.

The search was conducted on-line at the NCBI (<http://www.ncbi.nlm.nih.gov-/BLAST/>) and at the *A. thaliana* database (AtDB; Flanders, 1998) with the BLAST algorithm. Several sequences were found to be similar to the two cDNAs (see Section 10.3.2), but all except one did not correspond to either *PKT1* or *PKT2* transcripts. The exception was an EST derived from the cDNA clone 147C20T7 (Newman *et al.*, 1994). This EST was 367 residues-long and virtually identical to the 5' end of pCEN1-8. However, it failed to extend the known 5' terminal sequence of the *PKT1* transcript as it was 13 residues shorter than the sequence of pCEN1-8. The EST clone 147C20T7 was obtained from the *Arabidopsis* Biological Resource Center (Ohio State University, Columbus, Ohio) and plasmid DNA prepared. Its deriving from *PKT1* was confirmed by automated sequencing of the plasmid with the forward and reverse universal primers. At its 3' end the polyadenylation site was at the same position as in pCEN1-9, proving that both *PKT1* and *PKT2* transcripts may be polyadenylated at either of the two sites identified in the cDNA clones. More recently another EST, GenBank accession AI100014, was identified which derived from the 3' terminal region of a *PKT1* or *PKT2* transcript (see Section 10.3.2).

Figure 5.13 Sequences of the two transcription units (mP18 and mP19) at the tagged region.
Top: composite sequence of the first exon of pCEN1-8-type transcripts (mP18) is shown including the extensions obtained from the 5' RACE clones. The locations of the 5' ends of the 5' RACE sequences are indicated by the arrowheads. The remainder of the sequence, shared with mP19, is shown below.
Bottom: Composite sequence mP19 of the pCEN1-9-type transcripts including the 5' RACE extensions. The arrowheads indicate the 5' ends of the RACE sequences. The first exon of mP19 is shown in italics, the remainder of the sequence, shared with mP18, is written in normal typeface. At the 3' end the segment separating the two polyadenylation sites found in mP18 and mP19 transcripts is underlined.

Exon 1 of mP18

pTB-21, -24 pCEN1-5
∇ ∇ ∇ pCEN1-8
1 CATATTGTCTACTTTTAGATTTTGATATCTATTGGAACCTTACTAGCTC
51 ATTCATCCGTCGG

mP19

∇ pTB-6
1 TGAGACTGAGAACGATCGATCAAAAATCTCTCCCACGATAAAAATGAGGA

 ∇ pTB-3, -11 ∇ pTB-7 ∇ pTM-7
51 CCCTGAAGCTCATTGTACTTAACGTCTTCCCCTTCCAGAAAAATCAACC

 ∇ pCEN1-9
101 AAACCCATTTTCAGTTTCCCCGAAAGTTTTTGTCTTTCAATCACCATTAG

151 AATTCAGCTGCAACACAAGTGTTTGCAACAGAGAGAATCATGGAGAGAGC

 ∇ pCEN1-4 ∇ pCEN1-1
201 TATGGAAAGACAAAAGATATTGCTTCGTCATCTCAATCCAGTTTCTTCTT
251 CTAATTCTTCTCTTAAACATGAACCTTCTCTTCTGTCTCCTGTGAATTGT
301 GTTCTGAAGTTTCCCAATGGCTGCTTTTGGAGATGACATTGTGATTGT
351 AGCGGCATATCGTACCGCCATTTGCAAAGCGAGACGTGGAGGTTTCAAAG
401 ACACTCTTCCTGATGATCTTCTTGCTTCTGTTCTTAAGGCTGTAGTGGA
451 AGAACATCTTTGGATCCAAGTGAAGTTGGTGATATCGTTGTTGGTACCGT
501 GATAGCGCCTGGTTCTCAGAGAGCAATGGAGTGTAGAGTTGCAGCTTATT
551 TTGCTGGTTTTTCCTGACTCCGTGCCAGTTAGAACTGTCAATAGACAATGC
601 TCATCAGGACTACAAGCAGTTGCTGATGTTGCTGCTTCCATTAGAGCTGG
651 TTATTACGACATTGGTATTGGTGCTGGAGTGGAATCAATGTCAACTGATC
701 ATATTCCTGGAGGCGGCTTTTCATGGCTCTAATCCAAGAGCACAGGATTTTC
751 CCAAAGCCCGTGATTGTTTGCTTCCAATGGGAATTACTTCTGAAAACGT
801 TGCAGAAAGGTTCCGGTGTCACAAGAGAAGAGCAAGATATGGCTGCGGTGG
851 AGTCTCACAAACGCGCTGCAGCTGCAATCGCGTCTGGTAAACTCAAGGAT

901 GAAATCATTCCTGTTGCTACTAAGATTGTGGACCCTGAGACTAAAGCAGA
951 GAAGGCAATCGTCGTATCTGTTGATGACGGTGTACGTCCAAACTCAAACA
1001 TGGCAGATTTGGCAAAGCTGAAGACTGTCTTTAAACAGAACGGTTCCACC
1051 ACAGCTGGCAATGCTAGTCAGATCAGTGATGGTGCTGGAGCAGTACTGCT
1101 AATGAAGAGAAGTTTGGCTATGAAGAAGGGACTTCCCATTCTTGAGTAT
1151 TCAGGAGTTTTGCTGTTACTGGTGTGGAACCATCTGTAATGGGTATTGGT
1201 CCAGCTGTTGCCATTCCCGCTGCAACTAAGCTCGCAGGGCTCAACGTCAG
1251 CGATATTGATCTATTTCGAGATCAATGAGGCATTTGCATCTCAGTATGTGT
1301 ACTCTTGCAAGAAGTTAGAGCTGGATATGGAAAAGGTCAATGTTAATGGA
1351 GGAGCCATTGCTATTGGCCATCCCCTGGGTGCTACAGGAGCTCGATGTGT
1401 TGCACATTGTTGCACGAGATGAAGCGGAGAGGAAAAGATTGCCGCTTTG
1451 GAGTAATCTCAATGTGCATAGGCACTGGAATGGGAGCTGCAGCTGTTTTT
1501 GAGAGGGGAGACTCTGTTGATAACTTGTCCAACGCTCGTGTGGCTAACGG
1551 GGATAGTCATTAGAACATCGAAGAGAGCTTGAATAAGTAGAAGTAATGAT
1601 GCATTGAGTCTAATAAATATGATGCTTTAGCTCTTTCACATTGCTGAACA
1651 ATGAAAACTTTTGTCATTCTGAGTTTAAAATCAACTACTTTTCTCTG

5.6 Summary

The analysis of the inserts of the pCEN1-x clones (Section 4.4) at the sequence level was reported in this chapter. The presence of poly(dA) tails and the absence of intervening sequences (introns) in the sequences proved that they corresponded to cDNAs (Section 5.2.1). These cDNAs were found to correspond to two distinct but highly related transcripts identical throughout their length except for different exons at their 5' ends. The three clones pCEN1-1, -4 and -9 derived from the same transcription unit, that of a gene for 3-ketoacyl-CoA thiolase, *PKT2*, gene as described in Chapter 7. Its 5' limit in the cDNA clones was located 447 bp upstream from the E1.5 site. The orientation was opposite to the apparent orientation of the reporter gene in AtEN101.

The second transcript, derived from another gene for a 3-ketoacyl-CoA thiolase, *PKT1* (Chapter 7), was represented in pCEN1-5 and pCEN1-8. To determine if *PKT1* overlapped *PKT2* in the tagged region, or alternatively mapped elsewhere in the genome, Southern blot analysis was performed (Section 5.3.2). The results were consistent with the possible overlap between the two genes. This scenario was confirmed by sequencing a small genomic *EcoR* V fragment, E5.1E5.2, overlapping the T-DNA insertion site: the 5' end of the *PKT1* transcript derived from this region of the genome. It was separated from its exon 2 by a 966 bp-long intron that included the first known exon of *PKT2* (Section 5.3.4). No potential splice sites were found in the genomic segment separating the two first exons, in agreement with the hypothesis that the two types of transcripts corresponded to distinct genes (Section 5.3.5).

The relative positioning of the first exon of *PKT2* and of the T-DNA insertion site suggested that the expression of *gus* in AtEN101 might be related to *cis*-elements of the promoter of the native gene. To further approach the transcription start site of *PKT2* (and of *PKT1*) and hence define more precisely its promoter region, 5' RACE was performed (Section 5.4). Two established 5' RACE methods, SLIC (Section 5.4.2) and homopolymeric tailing-based (Section 5.4.4) were employed. Because of technical difficulties experienced with SLIC, a new 5' RACE method, LOPRACE, was also developed and applied (Section 5.4.3). With the same purpose, a search for *A. thaliana* ESTs derived from the *PKT1* and *PKT2* genes was conducted (Section 5.5). Two such ESTs corresponding to *PKT* transcripts were found. Together, the data permitted to extend the known sequences of the *PKT1* and *PKT2* transcripts at their 5' ends by 4 bp and 105 bp, respectively. This placed the T-DNA insertion site at a maximum distance of 337 bp upstream from the *PKT2* transcript.

Chapter 6

Complex organization of the T-DNA region in AtEN101

6.1 Introduction

The analysis of the *PKT1* and *PKT2* transcription units revealed how they map to the vicinity of the T-DNA insertion point in the genome of *Arabidopsis*, as described in the previous Chapter 5. It was there suggested that the observed expression of *gus* in AtEN101 might derive from the activity of a cryptic promoter (Section 5.4.1). It was further suggested that such a promoter might be related to the *cis*-elements regulating the expression of the *PKT2* gene. Co-localization of the promoter regions of both the *gus* reporter gene in AtEN101 and of *PKT2* would therefore be a pre-requisite of such scenario. Mapping the regions of the genome containing the 5' ends of these transcription units would enable the probing of this possibility and generally shed light on the nature of the promoter driving the expression of *gus* in AtEN101. For this reason, 5' RACE of fusion transcripts was performed as reported in this Chapter (Section 6.2).

Sequencing of several cDNA clones obtained by 5' RACE uncovered an unexpected complex organization of the T-DNA in the transgenic line. Both termini of the T-DNA in AtEN101 were found to consist of a left border region in a tail-to-tail orientation relative to each other. Furthermore, the evidence showed that fusion transcripts originated in the promoter region of *PKT1* and had the same orientation as that of the *PKT1* and *PKT2* genes.

Although no fusion transcripts with the opposite orientation were detected, their occurrence in AtEN101 could not be ruled out. To examine the potential expression of more than one *gus* gene in this line, detailed Southern and restriction enzyme analysis, and partial sequencing were performed to assess the number of reporter genes and their organization in the transgenic line (Section 6.3). The results clearly showed that at both T-DNA/native genome junctions a reporter gene was present. In addition, the restriction patterns of both *gus* genes were those expected if their structures were intact.

6.2 5' RACE of fusion transcripts

6.2.1 Application of the SLIC method

As indicated in Section 5.4, 5' RACE was first attempted in parallel for the *PKT* transcripts and the fusion transcripts with the SLIC method using a 5'-AmpliFINDER™ kit (Clontech, California). As suggested by the manufacturer of the kit, reverse transcription was performed

on 1.0 µg of poly(A)⁺ RNA of AtEN101. For a negative control, 0.5 µg of poly(A)⁺ RNA were submitted to the same procedures except that the reverse transcriptase was omitted from the reaction solution (Section 2.7.4.1). For mRNA isolation, total RNA was prepared from siliques of plants grown at 25°C under a day/night cycle (Section 2.6.2). The poly(A)⁺ RNA was then extracted from the total RNA with the aid of a suspension of poly(dT)-coated magnetic beads (Dynal), and purified (Section 2.6.4).

Reverse transcription of fusion transcripts was achieved with a primer designed from the *gus* gene sequence to ensure that it would not be located within a potential intron. The primer, prGUS-cDNA (TGACCGCATCGAAACGCAGCACG), mapped 298 bp downstream from the *gus* initiation codon and was oriented towards the left border. The cDNA obtained was purified and samples of it and of the negative control were submitted to PCR to confirm its synthesis. The primers used were prGUS-cDNA and prGUS-CONTROL (CCCTTATGTT-ACGTCCGTAGAAACCCC) which define a segment of 279 bp on the T-DNA sequence. Amplification was for 38 cycles and 20 µl aliquots were run on a gel. As can be seen in Figure 6.1, a PCR product of about 0.28 kb, the expected size, was present in the experimental sample but not in the negative control.

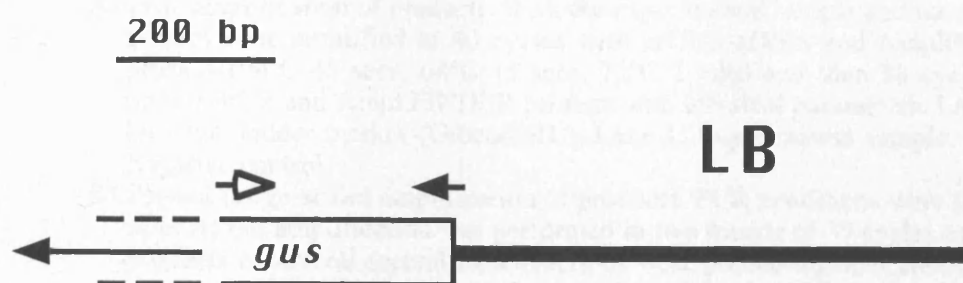
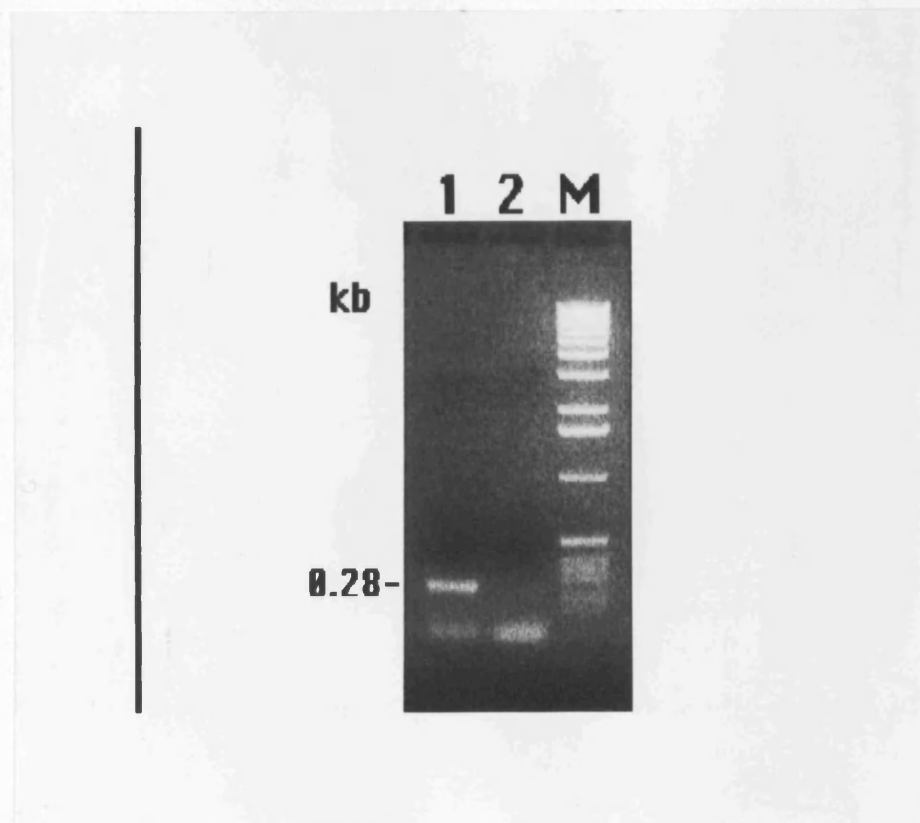
Having confirmed that its synthesis had occurred, the cDNA was ligated to the AmpliFINDER anchor. Given the high inefficiency of the ligation obtained with this method (see Section 5.4.1.1), the recommended protocol was modified by increasing the level of T4 RNA ligase activity and extending the period of incubation to 48 hours. PCR amplification ensued for 40 cycles with the supplied AmpliFINDER primer and prGUS-cDNA. No products could be seen after this first round of amplification. An aliquot of the reaction mixture was used to seed a second PCR reaction. This amplification was carried out with the anchor primer and prGUS-PCR (CAGGACGTAACATAAGGGACTGACCAC), a primer designed to overlap the sequence of prGUS-CONTROL and thus ensure its presence in the fusion transcripts. PCR products were observed at the end of 38 cycles of amplification (Fig. 6.2A). Bands of fragments of different sizes were observed. The most abundant were those of 0.20 kb and 0.35 kb. Other less abundant products of about 0.10 kb, 0.40 kb and 0.55 kb were also present. To obtain sufficient material for cloning the amplification procedure was scaled-up. PCR amplifications were performed consisting of two rounds of 37 cycles each. The products were pooled and concentrated by precipitation. An aliquot of the resuspended products was analyzed by agarose gel electrophoresis. The pattern of the pooled PCR products obtained was similar to that observed in the earlier experiment (Fig. 6.2B).

For cloning, the pooled PCR products were ligated into pCRTMII vector (Section 2.5.5.3). Following transformation of *E. coli*, the inserts in a number of clones obtained were sized by restriction enzyme analysis with *Eco*R I. The cloned fragments varied in size within the expected range from about 0.2 kb up to 0.4 kb (Fig. 6.3). These clones of SLIC products were designated pRF-x.

Figure 6.1 Confirming the synthesis of cDNA by reverse transcription of fusion transcripts. Reverse transcription of poly(A)⁺ RNA was performed with primer prGUS-cDNA as detailed in Section 2.7.4.1. A negative control lacking reverse transcriptase was set-up in parallel. Following purification of the samples, PCR amplification was performed for 38 cycles (94°C 45 secs, 66°C 35 secs, 72°C 1 min) with primers prGUS-cDNA and prGUS-PCR.

A) 20 µl samples of each PCR reaction were run in an agarose gel for resolution of the PCR products. Lane 1: Experimental sample. Lane 2: Negative control. Lane M: 1 kb DNA ladder size marker. The band of fragments of the expected size (0.28 kb) is indicated. The other fragments visible in both samples likely correspond to primer dimers.

B) Schematic representation of the T-DNA LB region of pΔ*gus*Bin19. The white arrowhead marks the location and direction of primer prGUS-cDNA. The filled arrowhead indicates the position of the primer prGUS-CONTROL. The orientation of the *gus* gene is indicated by the arrow.



6.2.5 Separation of 5' RACE products

Reverse transcription products were amplified with primers prGUS-PCR, prGUS-5, prGUS-6, prGUS-7 and prGUS-8. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

The products were separated by gel electrophoresis and stained with ethidium bromide. The products were separated by gel electrophoresis and stained with ethidium bromide.

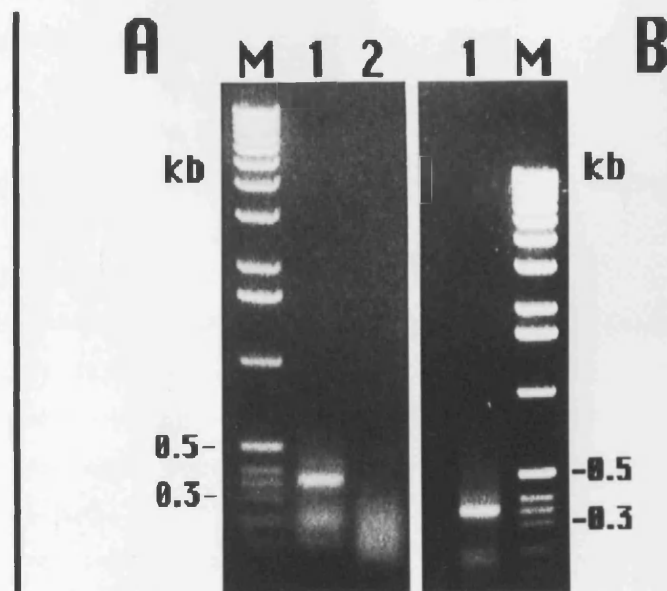


Figure 6.2 Visualization of 5' RACE products of fusion transcripts (SLIC method).

The cDNA obtained by reverse transcription was amplified in two rounds.

A) First amplification of products. Both the experimental sample and the negative control were amplified in 40 cycles with prGUS-cDNA and AmpliFINDER primers (94°C 45 secs, 64°C 45 secs, 72°C 2 min) and then 38 cycles with prGUS-PCR and AmpliFINDER primers with identical parameters. Lane M: 1 kb DNA ladder marker (Gibco/BRL); Lane 1: Experimental sample; Lane 2: Negative control.

B) Second (large-scale) amplification of products. PCR conditions were the same as in A) but amplification was performed in two rounds of 37 cycles each. The products of several second PCR reactions were pooled together and a sample analyzed by gel electrophoresis (Lane 1). Lane M: 1 kb DNA ladder marker.

6.2.2 Sequencing of SLIC products

Several clones with different insert sizes, including pRF-2, pRF-4, pRF-5, pRF-6, pRF-7 and pRF-8 (Fig. 6.3), were sequenced with the forward and/or reverse universal primers. The sequences of the inserts in these clones extended from the prGUS-PCR sequence to upstream points ranging from 164 bp (pRF-2) up to 329 bp (pRF-6) on the LB sequence and then were followed by the AmpliFINDER primer sequence. The length of the sequences fell shorter than that anticipated from the reported size of the fusion transcripts (see Section 4.1; Topping *et al.*, 1994). Also, no clones of the longer SLIC products were found. Nevertheless, the results established the presence of the sequenced segments in the fusion transcripts.

6.2.3 5' RACE by homopolymeric tailing

Because of the difficulties experienced in obtaining longer cDNA products with the SLIC method, the homopolymeric tailing method utilized with the *PKT1* and *PKT2* transcripts (Section 5.4.4) was also applied for 5' RACE of the fusion transcripts. Both procedures were performed in parallel using the same "5' RACE System for Rapid Amplification of cDNA Ends" kit from Gibco/BRL.

Total RNA was isolated from 8 day-old seedlings of AtEN101 grown in 1/2 MS₁₀ and reverse transcribed with the primer prGUS-PCR. After purification and 3' homopolymeric tailing with dC and terminal deoxynucleotidyl transferase, the cDNA was amplified. PCR was performed with prGUS-PCR and the supplied anchor primer for the poly(dC) tail. At the end of 32 cycles a second round of amplification was carried out for 25 cycles with the LB-267 and anchor primers. Upon agarose resolution of the PCR products two main bands of fragments of about 0.2 kb and 0.3 kb were visible (Fig. 6.4A). A faint band of about 0.7 kb fragments was also barely visible. To amplify the 0.7 kb fragments further, a small sample of the band was directly transferred from the gel to seed a third PCR amplification reaction as described in Section 2.7.4.3. This PCR was carried out for 32 cycles with the same primers and its products were resolved in an agarose gel (Fig. 6.4B). The small fragments and the 0.7 kb fragments were separately purified from the gels, ligated into pCRTMII vector and cloned in *E. coli*. The clones obtained were named pSFR-x for those derived from the smaller 5' RACE products, and pLFR-x for those obtained with the 0.7 kb fragments. Colony-PCR and/or restriction enzyme analysis were used to determine the insert sizes of the pSFR-x clones (Figs. 6.5) and pLFR-x clones (Fig. 6.6).

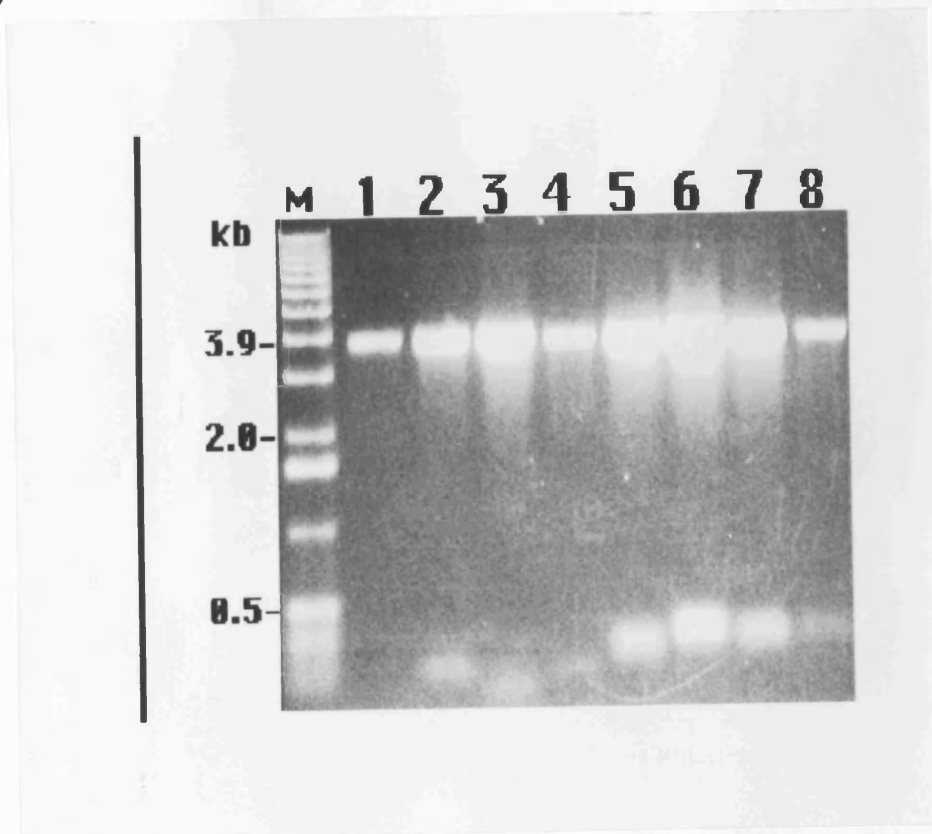
6.2.4 Fusion transcripts with unexpected 5' ends

The terminal sequences of the inserts in several pLFRs were manually determined including those of pLFR-3, pLFR-6, pLFR-7 and pLFR-9. All clones included at their termini the primer sequences, as expected, and a poly(dG) tract. Clone pLFR-6 was derived from an *A. thaliana* 18S rRNA, whereas pLFR-3 had no resemblance to any known sequence. The other

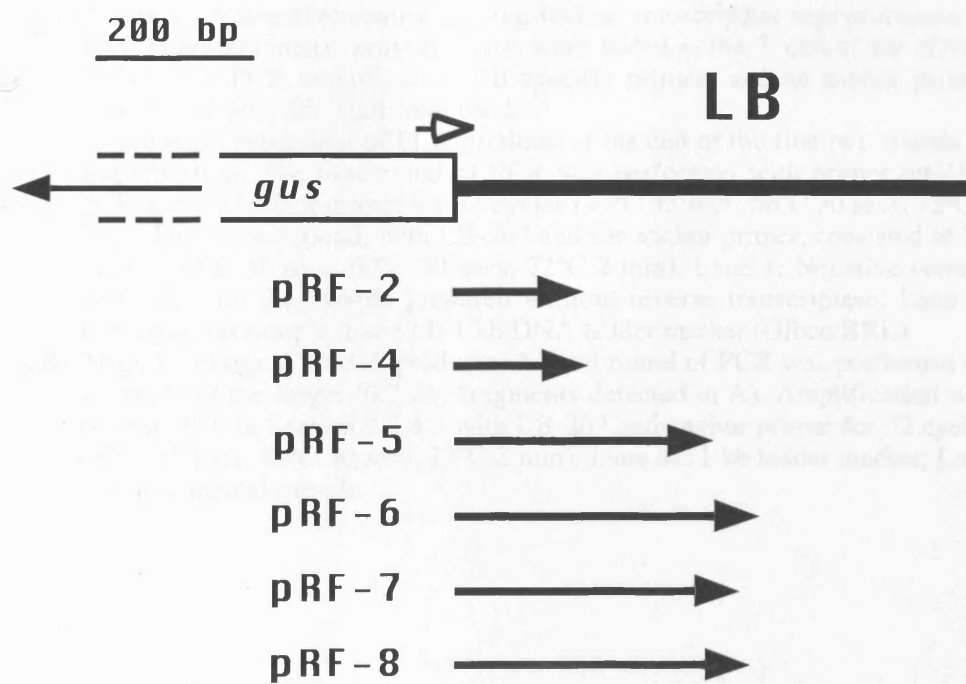
Figure 6.3 Sizing the 5' RACE inserts in pRF-x clones.

- A)** Plasmid DNA from various pRF-x clones was digested with *EcoR* I and the products were separated in an agarose gel. *EcoR* I sites are located on both polylinkers sequences flanking the insert in the clones and one is 4 bp downstream from the prGUS-PCR sequence in the LB region. Lane M: 1 kb DNA ladder marker (Gibco/BRL). Lanes 1-8: pRF-1 to pRF-8. The size of the larger pCRTMII-backbone is indicated (3.9 kb).
- B)** Map of the SLIC products present in the sequenced pRF clones. The size, location and orientation of the cDNA sequences are indicated by the arrows relative to the LB region of the T-DNA in p Δ gusBin19. The arrowhead marks the position and orientation of primer prGUS-PCR.

A)



B)



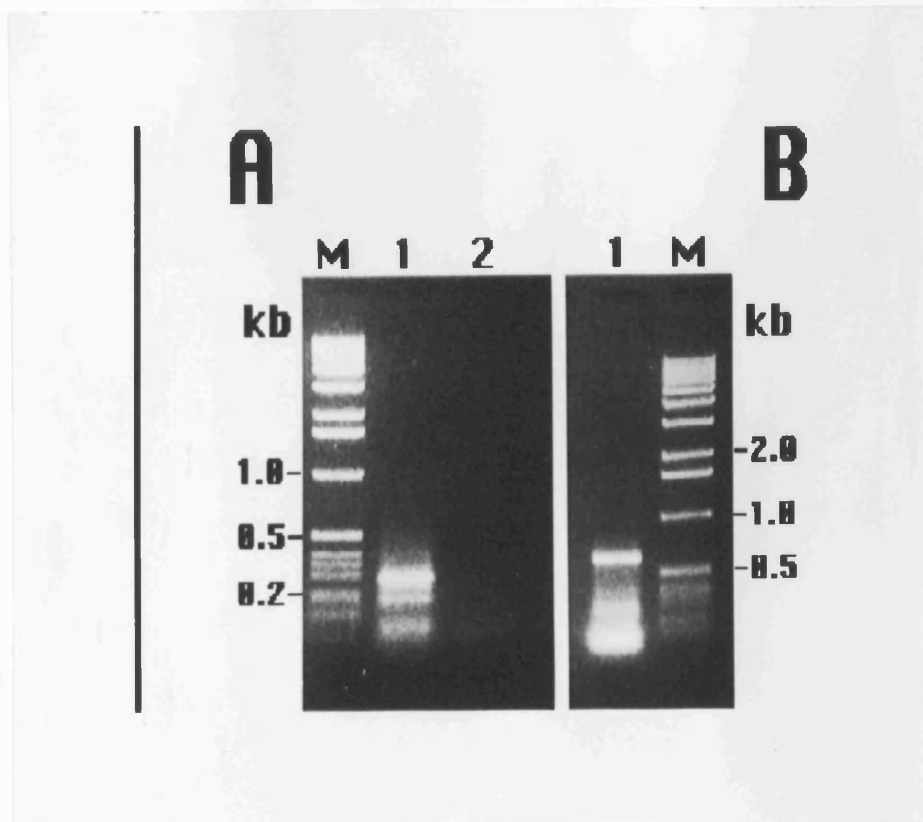


Figure 6.4 Amplification products of 5' RACE by homopolymeric tailing of fusion transcripts.

Reverse transcription was performed on total RNA of AtEN101 primed with prGUS-PCR. A negative control lacking reverse transcriptase was processed in parallel. Homopolymeric poly(dC) tails were added at the 3' end of the cDNA with TdT. For PCR amplification LB-specific primers and an anchor primer specific for the poly-(dC) tail were used.

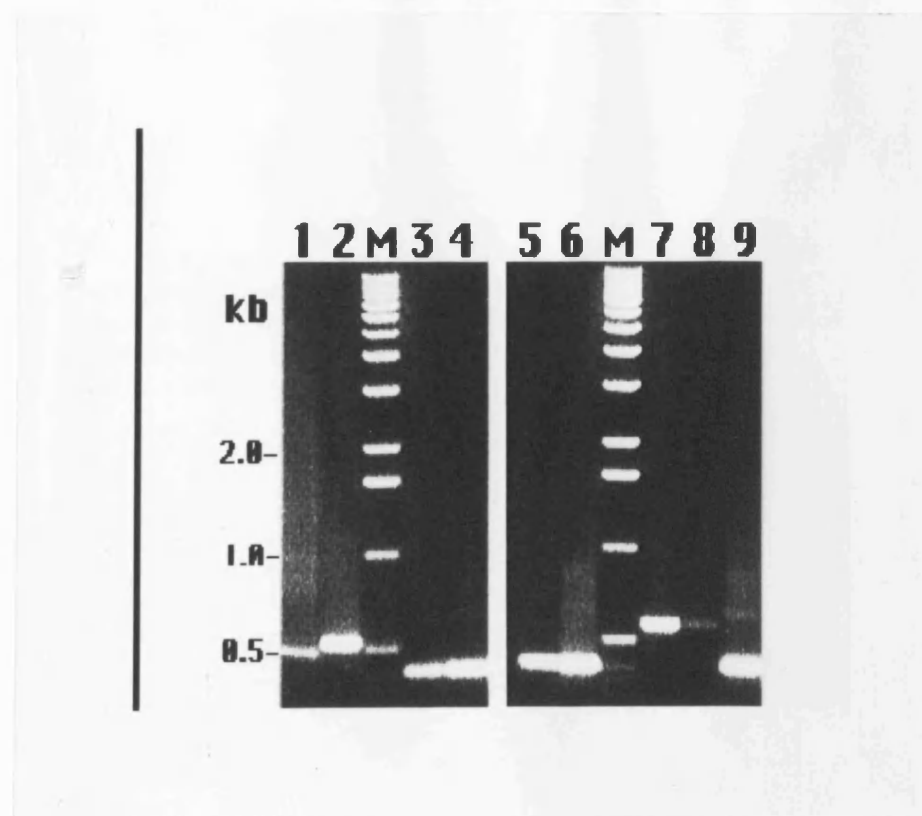
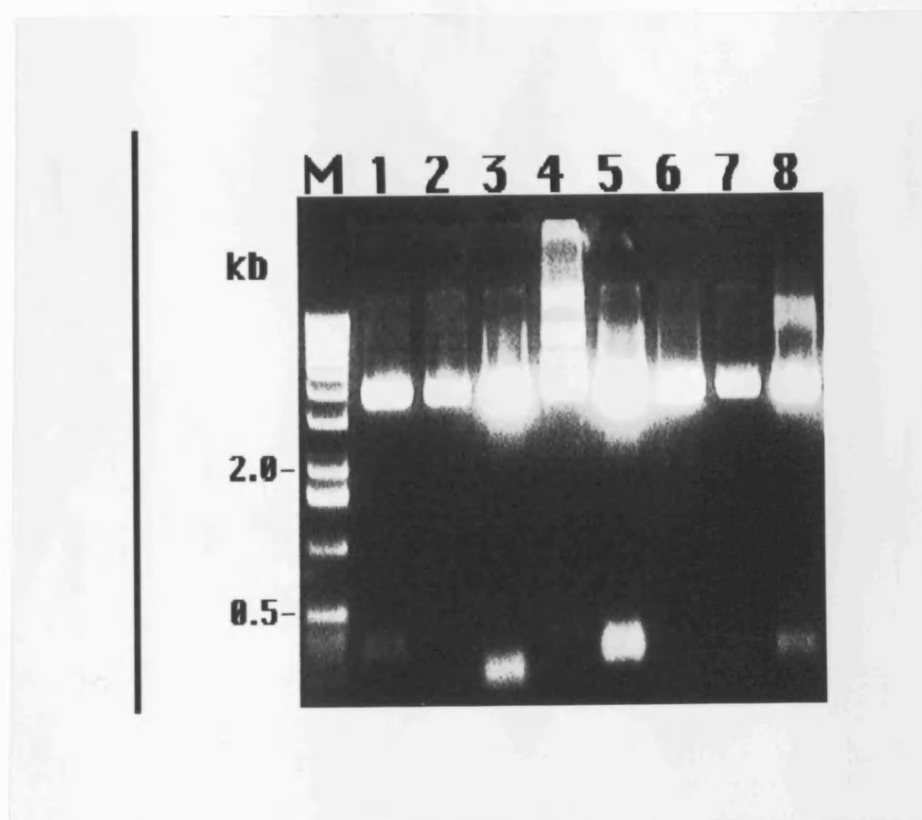
A) Agarose gel separation of PCR products at the end of the first two rounds of amplification. The first round of PCR was performed with primer prGUS-PCR and the anchor primer for 32 cycles (94°C 35 secs, 66°C 30 secs, 72°C 2 min). The second round, with LB-267 and the anchor primer, consisted of 25 cycles (94°C 35 secs, 60°C 30 secs, 72°C 2 min). Lane 1: Negative control performed on the sample prepared without reverse transcriptase. Lane 2: Experimental sample. Lane M: 1 kb DNA ladder marker (Gibco/BRL).

B) Amplified large 5' RACE products. A third round of PCR was performed on a sample of the larger (0.7 kb) fragments detected in A). Amplification was as described in Section 2.7.4.3 with LB-267 and anchor primer for 32 cycles (94°C 35 secs, 60°C 30 secs, 72°C 2 min). Lane M: 1 kb ladder marker; Lane 1: Experimental sample.

Figure 6.5 Sizing the inserts of pSFR-x clones.

Top: Sizing by restriction enzyme analysis. Several pSFR plasmids were digested with *Bst*X I to release their inserts, and the products resolved in an agarose gel. Lane M: 1 kb DNA ladder marker. Lanes 1-8: pSFR-1 to -8.

Bottom: Sizing by colony-PCR. The products obtained by colony-PCR (Section 2.5.5.4) with the forward and reverse universal primers of several pSFR-x clones were resolved in agarose gels. Lanes M: 1 kb DNA ladder marker. Lanes 1-9: pSFR-9 to -17.



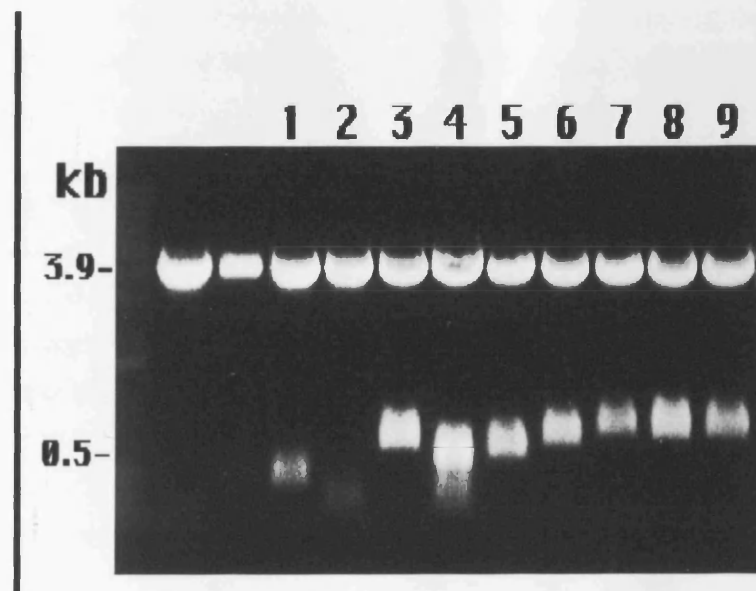


Figure 6.6 Sizing the inserts of pLFR-x clones by restriction enzyme analysis.

The plasmids were digested with *Bst*X I and the reaction products separated in an agarose gel. Lane M: 1 kb DNA ladder marker. Lanes 1-9: pLFR-1 to -9. The two non-labelled lanes corresponded to digests of pSFR clones.

two clones, pLFR-7 and -9, were virtually identical, albeit with different poly(dG) tracts of 12 and 14 residues respectively. Their sequences were longer than expected and derived solely from the binary vector. They included the LB region from -267 and extended into a position 540 bp upstream from the T-DNA in p Δ gusBin19 (Appendix 3).

The inserts of the pSFR-1, -4, -5, -8, -9, -10 and -15 were manually sequenced. The insert of pSFR-5 was flanked at one end by a poly(dG) tract and anchor primer sequence but the remainder lacked any similarity to known sequences. In all the other pSFR clones the inserts were flanked by a poly(dG) tract and anchor primer sequence at one end, and LB-267 at the other end. Their lengths varied from 235 bp (pSFR-9) to 337 bp (pSFR-4). From the LB-267 sequence, the identity to the LB region of the p Δ gusBin19 differed in its extension (Fig. 6.7).

In pSFR-1, -8, -9 and -10, the identity to the LB sequence terminated at the same position (Fig. 6.7). Surprisingly, in these four clones, the sequences upstream from the LB segment were derived from the first exon of *PKT1*. The identity to *PKT1* started at the 3' end of its first known exon and extended to different points upstream (Fig. 6.7A). The start of the longest *PKT1* segment in the clones, observed in pSFR-8, coincided with the known 5' limit of its transcripts (Section 5.4.4). Significantly, the point of transition between the *PKT1* and the LB segments coincided with the potential 3' splice site previously identified at position -101 of the LB sequence (see Section 4.2.6.3).

In clones pSFR-4 and pSFR-15 short segments of 47 bp and 28 bp, respectively, derived from the genomic fragment E5.1E5.2 were located downstream from the poly(dG) tails (Fig. 6.7B). The last 5 bp of these segments were shared with the T-DNA sequence which started at position -20.

The sequences of the 5' RACE clones demonstrated that the T-DNA in AtEN101 had a complex structure with, at least, one additional LB. The second LB was located downstream from exon 1 of *PKT1*, within the first intron of the gene, as indicated by the sequences of pSFR-4 and pSFR-15. The different structures of pSFR-1, -8, -9 and -10 are best interpreted as deriving from the splicing out of the junction segment downstream from exon 1 of *PKT1* up to the cryptic 3' splice site in the LB sequence (Fig. 6.7C). In conclusion, the sequence data suggested that the T-DNA in AtEN101 was flanked by two LB regions in a tail-to-tail orientation, as is shown in Figure 6.8. Hereafter, the junctions represented in pSFR-4 and pSFR-15, and that cloned in pIPCR2 will be referred as left and right junctions, respectively.

Figure 6.7 Structure of the 5' ends of fusion transcripts in AtEN101.

- A) Alignment of the sequences of the 5' RACE clones pSFR-x with homology to exon 1 of *PKT1* with the wildtype genome of *A. thaliana* and with the T-DNA LB region. Underlined are the matches to the 5' splice site consensus (GT) at the start of intron 1 of *PKT1* and the 3' splice site consensus (AG) in the LB sequence bordering the intron spliced out in the fusion transcripts. The anchor primer sequences upstream from the poly(dG) tracks are not shown. The E5.2 *EcoR* V site is shown underlined. The segments of the cDNA clones identical to regions of the LB region are shown in red.
- B) Alignment of the 5' RACE sequences in pSFR-4 and pSFR-15 with the wildtype genome at the T-DNA insertion region and with the LB sequence. Underlined in the wildtype genome sequence is the *EcoR* I site E1.5 (cf. Fig 4.11). In the wildtype sequence, the segment identical to the start of the sequence represented in pIPCR2 is shown in italics (Section 4.3). The segments of the cDNA clones identical to regions of the LB region are shown in red.
- C) Schematic representation of the structures of the pSFR-x clones and the predicted structure of the corresponding junction region in AtEN101. The two types of fusion transcripts are shown aligned with the proposed structure for the junction. Discontinuous lines denote non-sequenced segments (in the genome), or proposed spliced-out sequences (in the pSFR-8 type transcripts). The arrowhead indicates the position and direction of the LB-267 primer.

A)

pSFR-1 5'-GGGGGGGGTGTGATATCTATTGGAACCTTACTAGCTCATTATCCGTCGGT**TGAGACGGGAACAG**...

pSFR-9 5'-GGGGGGGGGGCTTACTAGCTCATTATCCGTCGGT**TGAGACGGGAACAG**...

pSFR-10 5'-GGGGGGGGGGGTTAGATTTTGATATCTATTGGAACCTTACTAGCTCATTATCCGTCGGT**TGAGACGGGAACAG**...

Wildtype genome ...CAACATTTGCATATTGTCTACTTTTAGATTTT**GATATC**TATTGGAACCTTACTAGCTCATTATCCGTCGGT**GT**AACTTTCTTGAC...

pSFR-8 5'-GGGGGGGGGGGCATATTGTCTACTTTTAGATTTTGATATCTATTGGAACCTTACTAGCTCATTATCCGTCGGT**TGAGACGGGCAACA**...

LB sequence (-90) ...CTTTTCACC**AG**TGAGACGGGCAACA...

B)

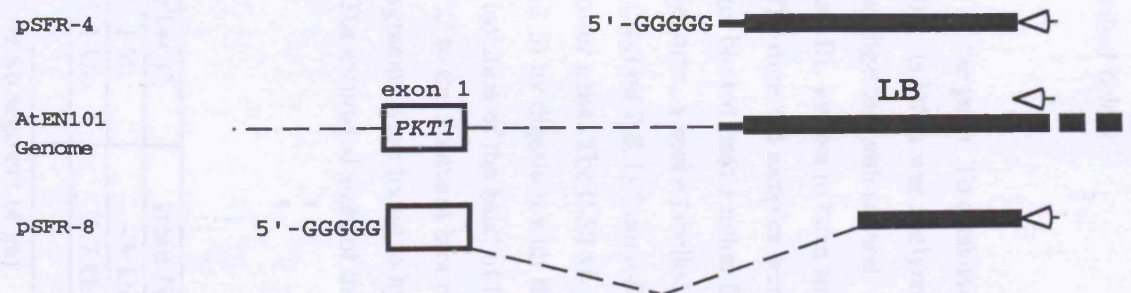
pSFR-4 5'-GGGGGGGGGGGGGAACAAAACAACCTTAAATCTATATCCGATATGAATTCTTGTGG**TAAACAAATTGACGCTT**...

Wildtype genome ...TTCTTACTAAACAAAACAACCTTAAATCTATATCCGATATGAATTCTTGTGG**TAAAGAATA**...

pSFR-15 5'-GGGGGGGGGGGGATATCCGATATGAATTCTTGTGG**TAAACAAATTGACGCTT**...

LB sequence (-1) **TGGCAGGATATATTGTGGTGTAAACAAATTGACGCTT**...

C)



6.3 Organization of T-DNA in AtEN101

6.3.1 Rationale

The organization of the T-DNA in AtEN101 suggested by the 5' RACE products of the fusion transcripts (Fig. 6.8) raised questions with regard to the source of *gus* expression in this line. Although the 5' RACE established that fusion transcripts originated from *PKT1*, it remained to be determined if they contained an intact *gus*-encoding sequence. Furthermore, the failure to detect fusion transcripts corresponding to the right junction constituted no evidence for their not occurring. Indeed, a cryptic promoter might be active at the right junction region under different physiological and/or developmental conditions. For these reasons, the *gus*-encoding potentials of the left and right junctions were assessed as described below.

6.3.2 Southern blot detection of *gus* genes in AtEN101

A pre-condition for detection of GUS activity is the intactness of the gene. To establish the potential number of intact *gus* genes in the genome of AtEN101, its DNA was analyzed by Southern blot hybridization. As a first step, genomic DNA was digested with several enzymes, including *Bam*H I, *Eco*R I, *Eco*R V, *Hinc* II and *Hind* III, known to have sites within the T-DNA region of pΔ*gus*Bin19 (Fig. 6.9B and C). The digested samples were then resolved by agarose gel electrophoresis (Fig. 6.9A), and the gel blotted onto a nylon filter (Section 2.5.3). To test the presence of *gus* sequences in the genome, a probe labelled with [α - 32 P]dCTP was prepared by random oligonucleotide priming (Section 2.8.1) from a *Bam*H I-*Eco*R V fragment containing the 5' terminal region of the reporter gene. The 0.58-kb long *gus* fragment was obtained from plasmid pPRGUS-6 (Section 2.1.2.5) by digestion with *Bam*H I and *Eco*R V, agarose gel separation of the reaction products, isolation of the band of interest and purification of the DNA. The probe was hybridized at 65°C to the Southern blot and the excess washed away as described in Section 2.8.3. Several fragments were found to hybridize to the probe, as shown in the autoradiogram in Figure 6.9A. The estimated sizes of the hybridizing fragments are listed in Table 6.1:

<i>Bam</i> H I	<i>Eco</i> R I	<i>Eco</i> R V	<i>Hinc</i> II	<i>Hind</i> III
> > 12 kb*	~9 kb	1.5 kb	4.1 kb	~9 kb
~9 kb*	-	1.3 kb	1.6 kb	5.7 kb
4.0 kb	-	-	-	-

Table 6.1 Genomic fragments hybridizing to the 5' terminal 0.58 kb segment of *gus*.

The sizes were estimated from comparison with the 1 kb DNA ladder marker run in parallel in the agarose gel shown in Figure 6.9. Asterisks mark fragments corresponding to fainter bands.

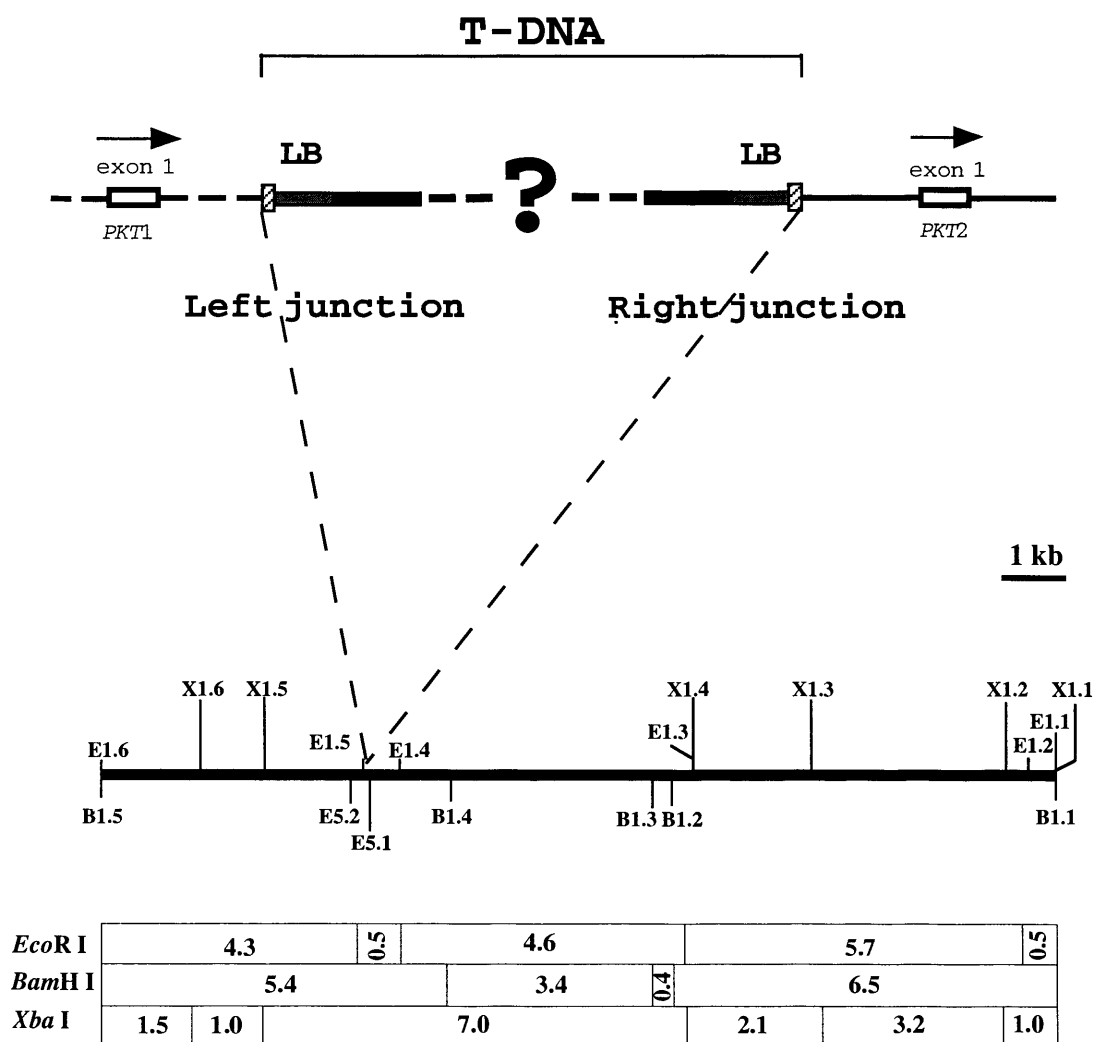


Figure 6.8 Proposed organization of the T-DNA in the genome of AtEN101.

Top: Organization of the T-DNA flanked by two LB regions in a tail-to-tail orientation. The arrangement at the left junction was suggested by the sequence data as described in the text. Discontinuous lines indicate undetermined sequences. The direction of transcription of *PKT1* and *PKT2* is shown by the arrows. Drawing not to scale.

Bottom: The point of insertion of the T-DNA in the restriction map of the wildtype genome is shown. The restriction map is derived from that in Figure 4.11 after inversion of the latter to ascribe the conventional representation of orientation to the *PKT1*, *PKT2* and fusion transcripts.

Figure 6.9 Southern blot detection of 5' *gus* regions in AtEN101.

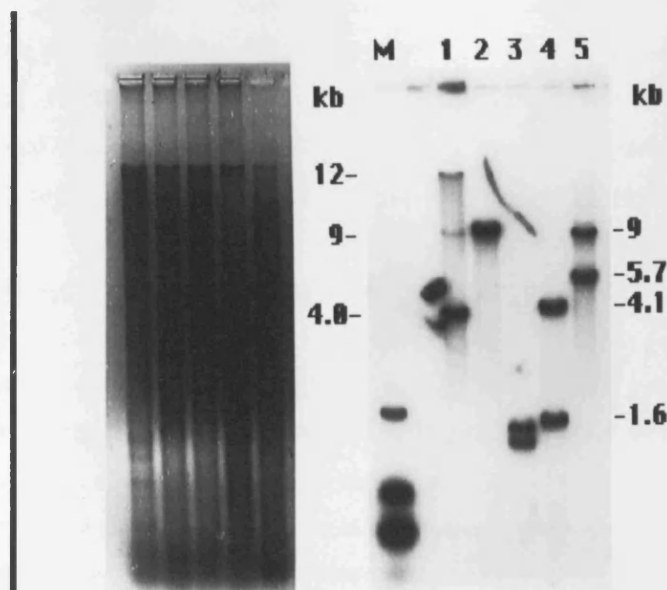
A) Aliquots of total DNA of AtEN101 were digested with *Bam*H I, *Eco*R I, *Eco*R V, *Hinc* II and *Hind* III. Following resolution by electrophoresis in a 0.7% agarose gel (Left) a Southern blot was prepared. The blot was hybridized to a probe prepared from the 0.58 kb 5' terminal region of *gus* by random oligonucleotide priming (Section 2.8.1). Hybridization and washes were at 65°C. An autoradiogram of the hybridized blot is shown (Right). Lane M: 1 kb DNA ladder marker; Lane 1: *Bam*H I; Lane 2: *Eco*R I; Lane 3: *Eco*R V; Lane 4: *Hinc* II; Lane 5: *Hind* III.

B) Predicted structure of the left T-DNA junction. The map assumes that an intact *gus* gene and associated *nos* terminator are present downstream from the LB region.

C) Predicted structure of the right T-DNA junction. The map assumes that an intact *gus* gene and associated *nos* terminator are present upstream from the LB region.

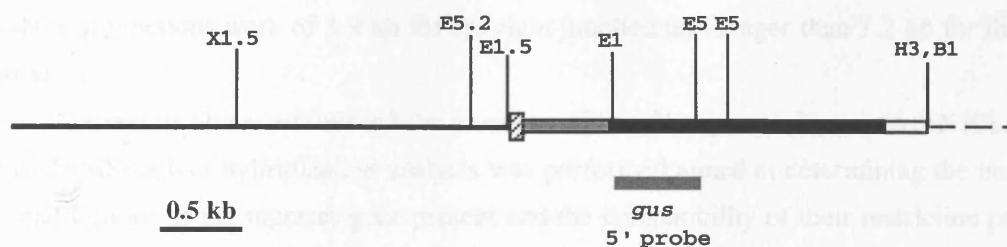
In B) and C) the restriction enzymes sites for *Bam*H I (B1), *Eco*R I (E1), *Eco*R V (E5), *Hinc* II (Hc), *Hind* III (H3) and *Xba* I (X1) are indicated. Sites located in the native genome region are identified according to the map in Figure 6.8. The localization of the *gus* fragment used to prepare the probe is also indicated.

A)



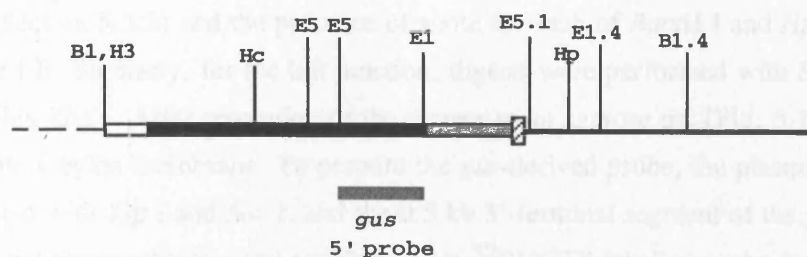
B)

Left junction



C)

Right junction



The results demonstrated that there were two distinct fragments in the genome of AtEN101 similar in sequence to the *gus* probe. Indeed, occurrence of two *EcoR* V fragments hybridizing to the probe is good evidence of it as the activity of this enzyme is not affected by cytosine methylation. Therefore the bands could not have resulted from partial digestion due to methylation interference. Furthermore, the 1.3 kb and 1.5 kb *EcoR* V fragments have the sizes expected from the predicted structures at the right and left junctions, respectively (Fig. 6.9B and C). The *Hind* III digest was also anticipated to have been nearly complete as the enzyme recognition sequence (AAGCTT) is not a target for cytosine methylation in plants. As a result, the two bands of *Hind* III fragments hybridizing to the *gus* probe with comparable strength ought to have derived from distinct reporter genes.

The single band of *EcoR* I fragments (~9 kb) is compatible with the presence of two *gus* genes in the T-DNA as it may correspond to two large fragments similar in size or to a single fragment that encompasses both genes. In this regard, it is noteworthy that the T-DNA of pΔ*gus*Bin19 contains a unique *EcoR* I site located at the 5' terminus of *gus*. The pattern of hybridization obtained with the *Bam*H I digest was incongruent in that a single strong band of 4.0 kb fragments and two weaker bands of much larger fragments were observed. This likely resulted from poor digestion at one of the sites flanking the 9 kb fragment, perhaps due to methylation interference, and deficient transfer of large DNA segments to the membrane during blotting. This scenario was in agreement with the predicted structures at the junctions as deduced from the known sequence of the T-DNA in pΔ*gus*Bin19 and the sequence and restriction data available: the estimated lengths for the *Bam*H I fragments containing the hypothetical junctions were of 3.9 kb for the right junction and longer than 7.2 kb for the left junction.

The results above established the presence of two 5' regions of *gus* in AtEN101. Further Southern blot hybridization analysis was performed aimed at determining the number of 3' end regions of the reporter gene present and the compatibility of their restriction patterns with those expected from the predicted structures of the T-DNA junction regions. Restriction maps for the predicted junctions are shown in Figures 6.10B and C where the *gus* genes are presumed to be intact. To observe changes in the restriction pattern at the right junction digests were carried out with *Hind* III, *Hind* III plus *Hpa* I, and *Hind* III plus *Bam*H I. The enzymes were chosen taking in consideration the lack of *Hind* III sites in B1.4B1.5 genomic segment (Section 5.3.3) and the presence of a site for each of *Bam*H I and *Hpa* I downstream from right LB. Similarly, for the left junction, digests were performed with *Bam*H I, and *Bam*H I plus *Xba* I. After resolution of the digests in an agarose gel (Fig. 6.10A), this was blotted onto a nylon membrane. To prepare the *gus*-derived probe, the plasmid pPRGUS-6 was digested with *Ssp* I and *Sac* I, and the 0.5 kb 3'-terminal segment of the gene was isolated following gel electrophoresis and purified. A [α -³²P]dCTP-labelled probe was then obtained from the *gus* fragment by random oligonucleotide priming. The probe was hybridized to the

Southern blot at 65°C, and, after washing the excess probe, an X-ray film was exposed to the blot. The autoradiographic results are shown in Figure 6.10A and the sizes of the fragments hybridizing to the probe are listed in Table 6.2.

<i>Hind</i> III	<i>Hind</i> III + <i>Hpa</i> I	<i>Hind</i> III + <i>Bam</i> H I	<i>Bam</i> H I	<i>Bam</i> H I + <i>Xba</i> I
9 kb	9 kb	9 kb	> > 12 kb	9 kb*
5.7 kb	3.1 kb	3.9 kb	9 kb*	4.7 kb
-	-	-	3.9 kb	3.8 kb

Table 6.2 Genomic fragments of AtEN101 hybridizing to the 3' region of *gus*.

The sizes of the fragments hybridizing to *Ssp* I-*Sac* I 0.5 kb segment of *gus* were estimated from the gel in Figure 6.10A. Asterisks mark very weakly hybridizing bands.

The hybridization patterns showed the presence of two 3' end segments of *gus* in AtEN101. In addition, the smaller hybridizing fragments visible in the *Hind* III digests had the sizes expected if an intact *gus* gene was present at the right junction (Fig. 6.10C). The *Bam*H I pattern of hybridization was similar to that observed with the probe from the 5' end of *gus* except that the larger band had a strength comparable to that of the 3.9 kb fragments. Also, digestion with *Bam*H I and *Xba* I reduced the size of the larger *Bam*H I fragments to 4.7 kb, in good agreement with the presence of an intact *gus* gene at the left junction (Fig. 6.10B). However, the small reduction in the size of the *Bam*H I 3.9 kb fragments to 3.8 kb by digestion with *Xba* I was unexpected. The most likely explanation for this observation is the presence of an *Xba* I site at the right junction region at a point anticipated to correspond to the *nos* terminator. As a consequence, it was uncertain if the *gus* gene in the right junction region was intact at its 3' end.

6.3.3 Restriction enzyme analysis of the T-DNA junctions

6.3.3.1 PCR amplification of the left and right junctions

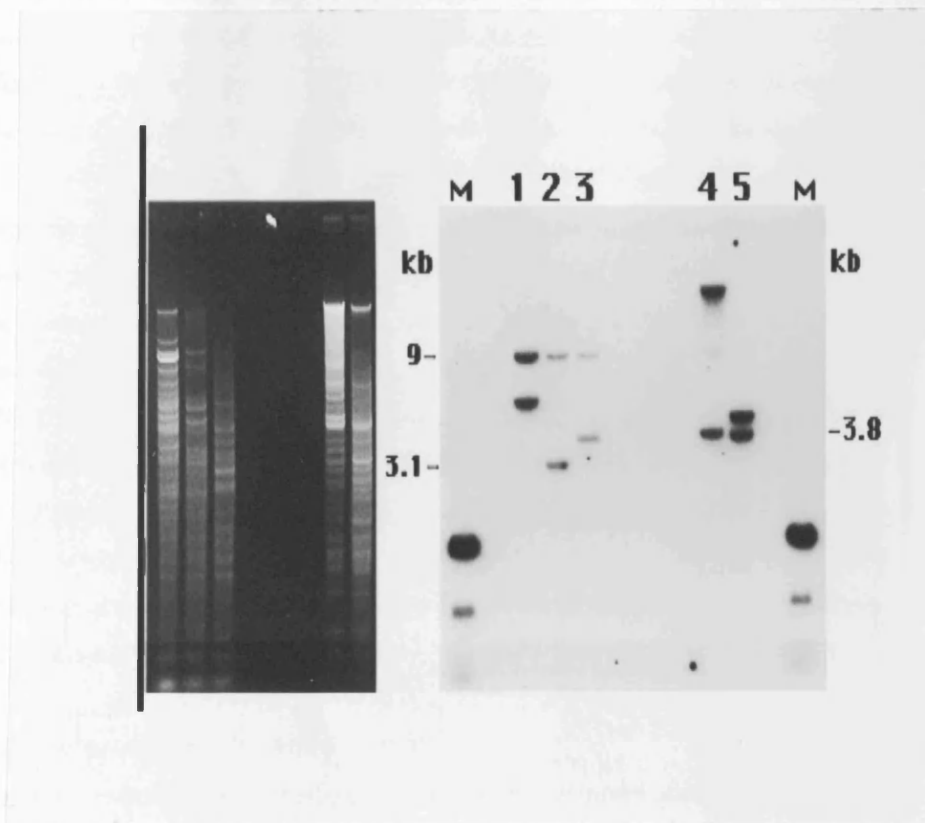
The Southern blot analysis demonstrated the presence of two *gus* genes in AtEN101 but was insufficient to determine if their structures were intact. To obtain more detailed information in this regard, the left and right junctions were subjected to detailed restriction enzyme analysis.

With the purpose of obtaining the necessary material for the analysis, the junctions were amplified by PCR from the genome of AtEN101 (Section 2.5.5.2). For the amplification a primer, prGUS-3' (GCCGCCGACTTCGGTTTGCGGTCG), was designed that annealed 67 bp upstream from the stop codon in the *gus*-coding sequence. The primer prGUS-3' was used

Figure 6.10 Southern blot detection of 3' regions of *gus* in AtEN101.

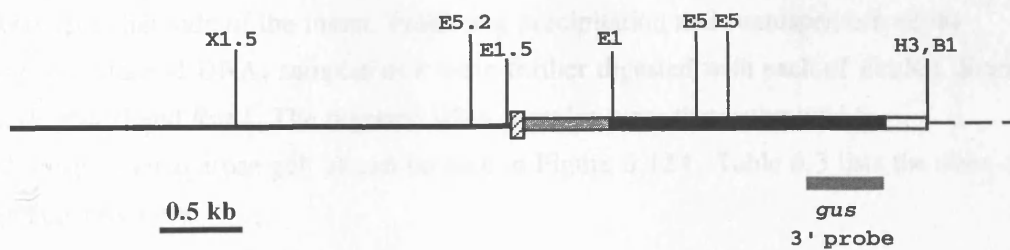
- A) Various single enzyme and double restriction enzyme digests of total DNA of AtEN101 were prepared. The digests were electrophoretically resolved in a 0.7% agarose gel (Left) and then blotted onto a nylon membrane. The Southern blot was hybridized at 65°C with a probe prepared from the 0.5 kb-long 3' terminal region of *gus* by random oligonucleotide priming (Section 2.8.1). Washing was at 65°C. An autoradiogram of the hybridized blot is shown (Right). Lane 1: *Hind* III; Lane 2: *Hind* III + *Hpa* I; Lane 3: *Hind* III + *Bam*H I; Lane 4: *Bam*H I; Lane 5: *Bam*H I + *Xba* I.
- B) Predicted structure of the right T-DNA junction. The map assumes that an intact *gus* gene and associated *nos* terminator are present upstream from the LB region.
- C) Predicted structure of the right T-DNA junction. The map assumes that an intact *gus* gene and associated *nos* terminator are present downstream from the LB region.
- In B) and C) the restriction enzymes sites for *Bam*H I (B1), *Eco*R I (E1), *Eco*R V (E5), *Hind* III (H3), *Hpa* I (H) and *Xba* I (X1) are indicated. Sites located in the native genome region are identified according to the map in Figure 6.8. The location of the *gus* fragment used to prepare the probe is also indicated.

A)



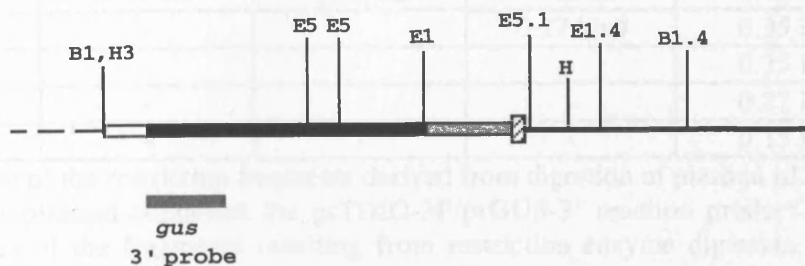
B)

Left junction



C)

Right junction



in conjunction with the primer prTHIO-3F (CTCTTACTAGCTCATTTCATCCG) for amplification of the left junction, and with prEN1-P3 for amplification of the right junction. PCR was performed with an ExpandTM PCR system (Boehringer Mannheim) under the conditions described in Section 2.5.5.2, together with negative controls. The annealing temperatures were 66°C and 57°C for the primer pairs prEN1-P3/prGUS-3' and prTHIO-3F/-prGUS-3', respectively. The products of the PCR reactions were resolved in an agarose gel (Fig. 6.11). Single products of 3.3 kb and 2.7 kb were obtained for the right junction and the left junction reactions, respectively. The sizes of these products were similar to those expected if the junctions had intact *gus* genes.

The PCR reactions were scaled-up and repeated under identical conditions to obtain sufficient material for the subsequent manipulations. The products were purified from the PCR reaction solution with the aid of a "High pure PCR product purification" kit (Boehringer Mannheim). Following precipitation and resuspension in TE buffer, aliquots were run in an agarose gel to confirm that the expected products had been obtained and were not damaged during the purification steps (data not shown).

6.3.3.2 Restriction analysis of the left junction

The 2.7 kb fragment corresponding to the left junction was ligated into a T/A-type vector pGEMTM-T (Promega), and clones of it were obtained by transformation of *E. coli* with the ligation products. Three plasmids containing the junction fragment were obtained. One of these, pJ2-2, was subjected to digestion with *Apa* I and *Pst* I, enzymes that cut the vector polylinkers on either side of the insert. Following precipitation and resuspension of the digested pJ2-2 plasmid DNA, samples of it were further digested with each of *Eco*R I, *Eco*R V, *Hinc* II, *Pvu* II and *Rsa* I. The digested DNA samples were then submitted to electrophoresis in an agarose gel, as can be seen in Figure 6.12A. Table 6.3 lists the sizes of the fragments observed.

<i>Eco</i> R I	<i>Eco</i> R V	<i>Hinc</i> II	<i>Pvu</i> II	<i>Rsa</i> I
3.0 kb #	3.0 kb #	3.0 kb #	2.8 kb #	1.8 kb #
1.8 kb	1.6 kb	1.6 kb	1.9 kb	1.2 kb #
0.7 kb	1.0 kb	0.65 kb	0.48 kb	0.95 kb
0.27 kb	0.30 kb	0.55 kb	0.42 kb	0.60 kb
-	-	-	0.27 kb #	0.35 kb
-	-	-	-	0.35 kb
-	-	-	-	0.22 kb
-	-	-	-	0.15 kb

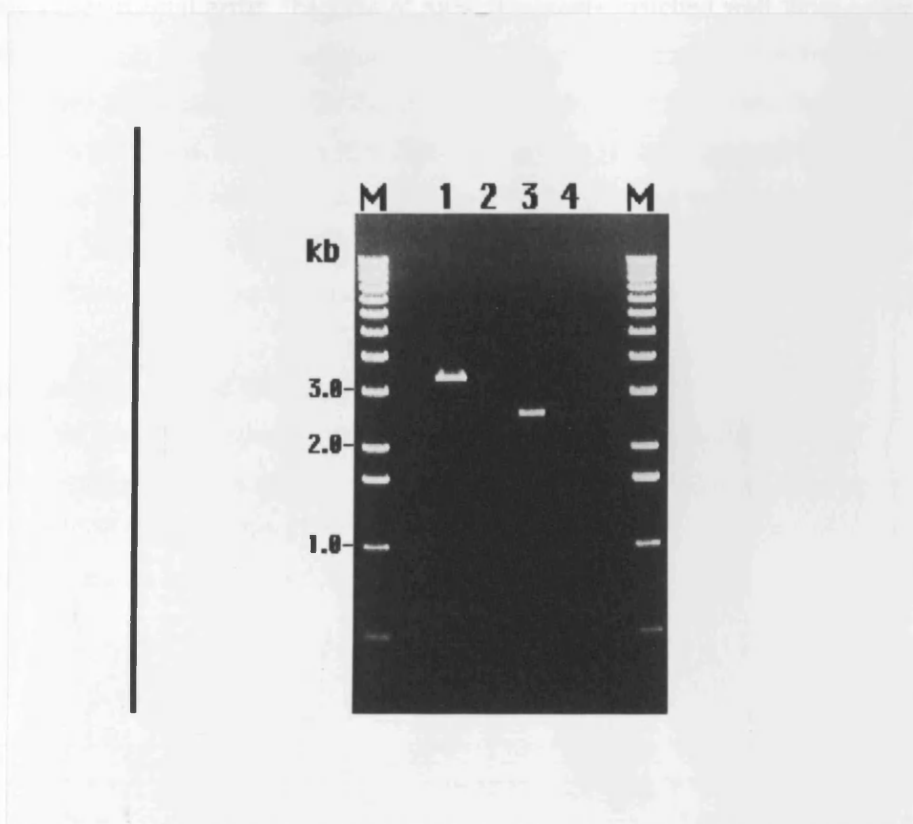
Table 6.3 Sizes of the restriction fragments derived from digestion of plasmid pJ2-2.

The plasmid contained the prTHIO-3F/prGUS-3' reaction product. The sizes of the fragments resulting from restriction enzyme digestion were estimated from the gel in Figure 6.12A. Those fragments derived from the cloning vector pGEMTM-T are marked with "#".

Figure 6.11 PCR amplification of the T-DNA junctions of AtEN101.

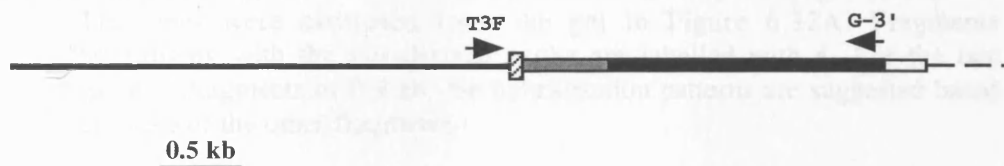
- A)** The left and right junctions were amplified from total DNA of AtEN101 with an Expand[™] PCR system (Boehringer-Mannheim) and the primer pairs prTHIO-3F/prGUS-3' and prEN1-P3/prGUS-3', respectively. Negative controls for each reaction were similarly prepared that lacked the genomic DNA. PCR conditions were as described in Section 2.5.5.2. PCR conditions were as described in Section 2.5.5.2. Annealing temperatures were 66°C and 57°C for the primer pairs prEN1-P3/prGUS-3' and prTHIO-3F/prGUS-3', respectively. The products of the reactions were resolved in the 0.8% agarose gel shown. Lane M: 1 kb DNA ladder marker. Lane 1: Experimental right junction amplification; Lane 2: Negative control (prEN1-P3/prGUS-3'); Lane 3: Experimental left junction control; Lane 4: Negative control (prTHIO-3F/-prGUS-3').
- B)** Schematic representation of the expected structure of the left junction region. Indicated are the relative locations of the annealing sites for primers prTHIO-3F and prGUS-3', and their orientations. The *gus* gene is shown as a thick solid bar.
- C)** Schematic representation of the expected structure of the right junction region. Indicated are the relative locations of the annealing sites for primers prEN1-P3 and prGUS-3', and their orientations. The *gus* gene is shown as a thick solid bar.

A)



B)

Left junction



C)

Right junction



Within experimental error, the sizes of most fragments matched well those expected for a structure of the left junction containing an intact *gus* gene as predicted in Figure 6.10B. The only exceptions were found with the *EcoR* V digests where all fragments seemed to be slightly larger than predicted, including that derived from the vector sequence. This latter observation and the requirement for the overall length of the plasmid to be constant indicates that the apparently larger than expected sizes of the *EcoR* V fragments were a gel artifact. It possibly resulted from a retarding effect of the restriction enzyme buffer used in the digestion.

6.3.3.3 Restriction analysis of the right junction

The amplified right junction fragments were directly submitted to restriction analysis following their purification from the PCR reaction solution. Several aliquots were digested with *EcoR* I, *EcoR* V, *Hinc* II, *Ssp* I and *Rsa* I. The products of digestion were then resolved in an agarose gel, shown in Figure 6.13. The sizes of the fragments obtained are listed in Table 6.4.

<i>EcoR</i> I	<i>EcoR</i> V	<i>Hinc</i> II	<i>Ssp</i> I	<i>Rsa</i> I
1.7 kb #	1.3 kb #	1.5 kb #	0.95 kb #	2.9 kb #
1.2 kb	0.95 kb #	0.70 kb	0.60 kb #	0.48 kb #
0.45 kb	0.94 kb	0.60 kb #	0.47 kb	-
-	0.22 kb	0.50 kb #	0.47 kb	-
-	-	-	0.33 kb #	-
-	-	-	0.20 kb #	-
-	-	-	0.15 kb #	-

Table 6.4 Restriction fragments obtained by digestion of the amplified right border.

The sizes were estimated from the gel in Figure 6.12A. Fragments hybridizing with the *gus*-derived probe are labelled with #. For the two *EcoR* V fragments of 0.9 kb, the hybridization patterns are suggested based on those of the other fragments.

The lengths of the observed fragments were found to be in good agreement with the expected restriction pattern for a right junction containing an intact *gus* gene (Fig. 6.13). It could then be concluded that a reporter gene which had not suffered any major structural changes was present upstream from the right LB.

To confirm the origin of the various fragments presumed to derive from *gus*, a Southern blot was prepared from the gel and hybridized with a probe derived from the reporter gene. The [α - 32 P]dCTP-labelled probe was synthesized from the entire *gus* gene in plasmid pPRGUS-6 following isolation and purification of the corresponding *Sac* I fragment (Section 2.4.6). Hybridization of the blot and washes were at conducted at 68°C (Section

Figure 6.12 Structure of the T-DNA left junction in AtEN101.

Top: Electrophoretic resolution of restriction enzyme digests of plasmid pJ2-2 (pre-digested with *Apa* I and *Pst* I). Lanes M: 1 kb DNA ladder marker. Lane 1: *EcoR* I; Lane 2: *EcoR* V; Lane 3: *Hinc* II; Lane 4: *Pvu* II; Lane 5: *Rsa* I. *Pvu* II digests the cloning vector sequence outside its polylinker regions generating two fragments of 0.27 kb and 2.8 kb visible in the gel. Similarly, *Rsa* I digests the vector into a 1.8 kb and a 1.2 kb fragments.

Bottom: Restriction patterns expected for the insert of pJ2-2 if, with exception of the junction segment sequenced in pSFR-4 and pSFR-15, the remainder of the sequences between prTHIO-3F and prGUS-3' were intact. The restriction enzymes sites are marked: *EcoR* I (E1), *EcoR* V (E5), *Hinc* II (Hc), *Pvu* II (P) and *Rsa* I (R). The sizes in bp of the fragments are also indicated.

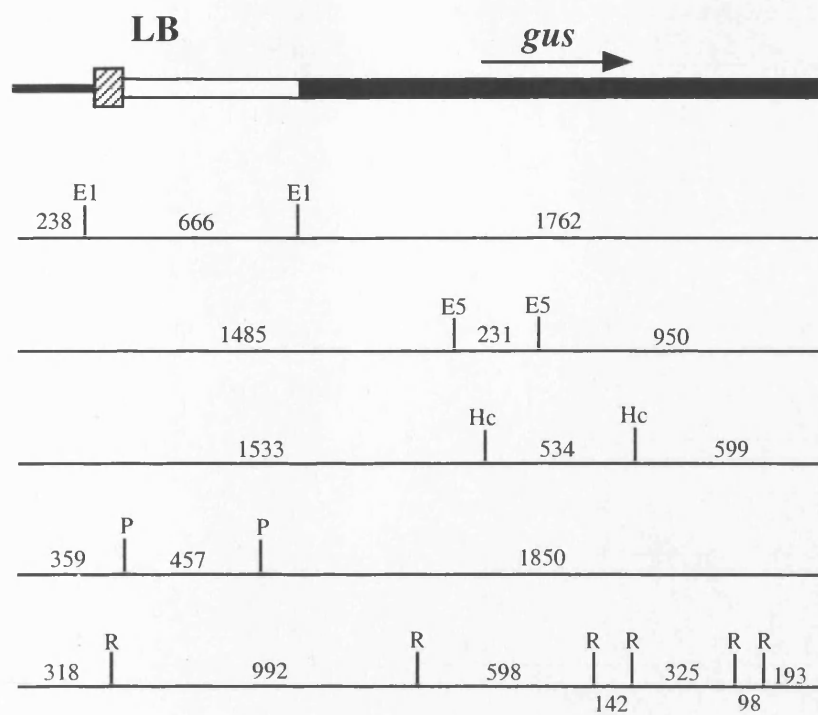
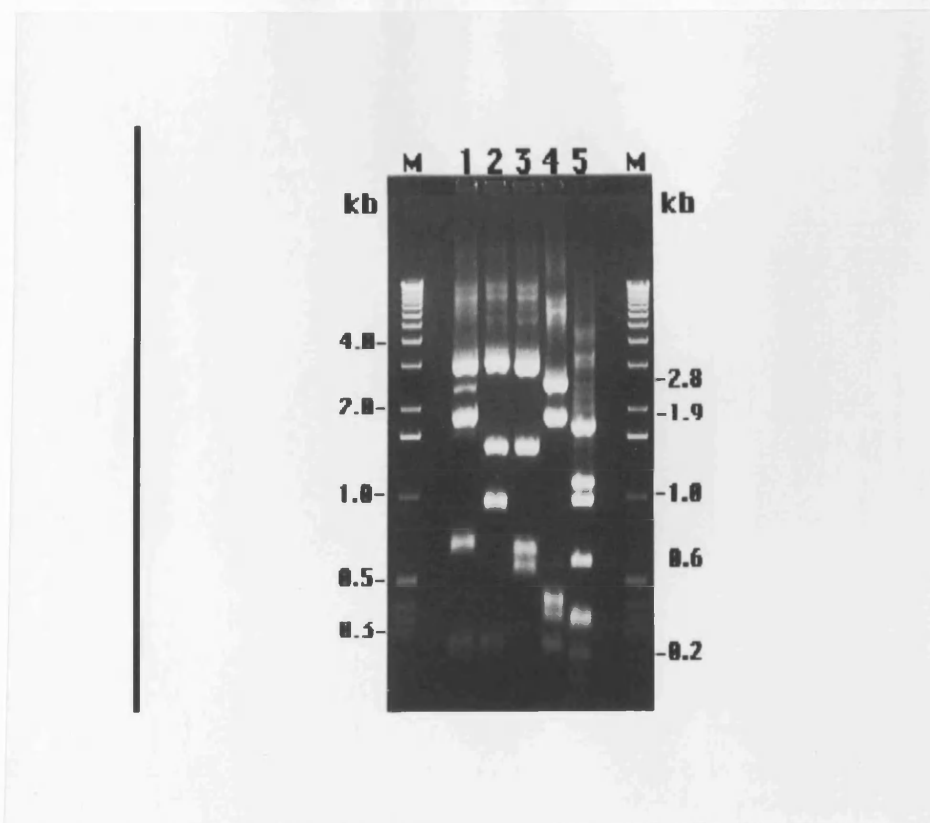


Figure 6.13 Structure of the T-DNA right junction in AtEN101.

Top: The restriction enzyme digests of the PCR amplified junction were separated in the agarose gel shown (Left). To enable the visualization of small and larger fragments two different images of the gel, differing in exposure, are shown. A autoradiograph of a Southern blot of the gel hybridized to a probe derived from the entire *gus* gene can also be seen (Right). Hybridization and washes were at 68°C.

Bottom: Restriction patterns predicted for the PCR product presuming that the T-DNA sequence downstream from prGUS-3' was intact. The restriction enzymes sites are marked: *EcoR* I (E1), *EcoR* V (E5), *Hinc* II (Hc), *Rsa* I (R) and *Ssp* I (S). The sizes in bp of the fragments are also indicated.

2.5.3. An additional set of bands (Fig. 2.5.3) was observed when a complete

chromosome was inserted, the fragments of DNA that were supplied and shown

to be

the right

fragments

the right

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

fragments

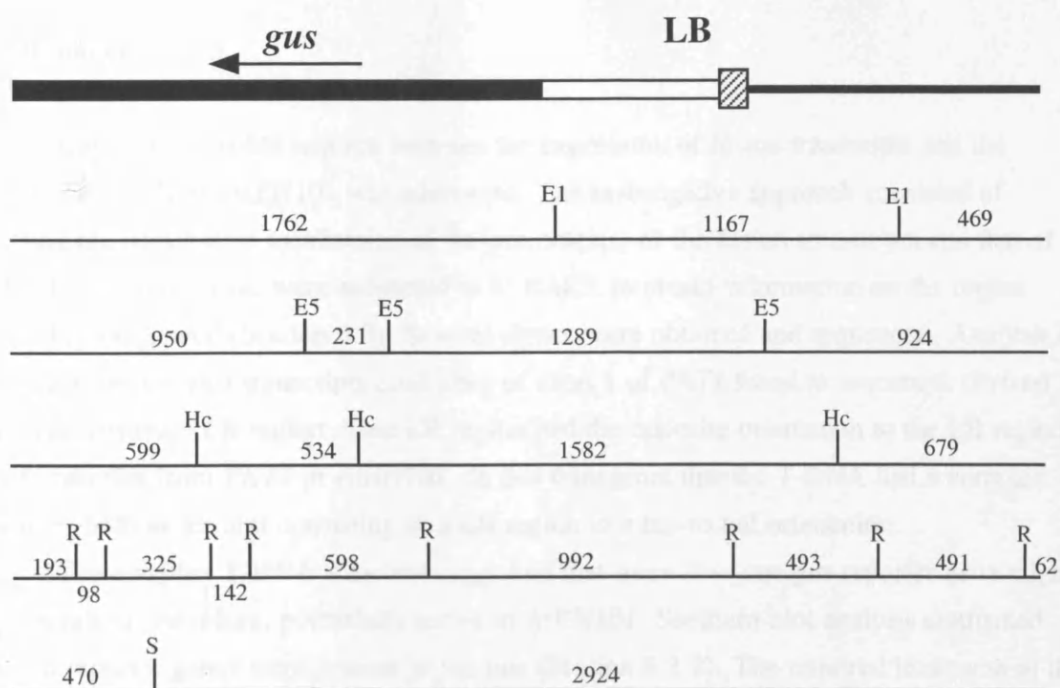
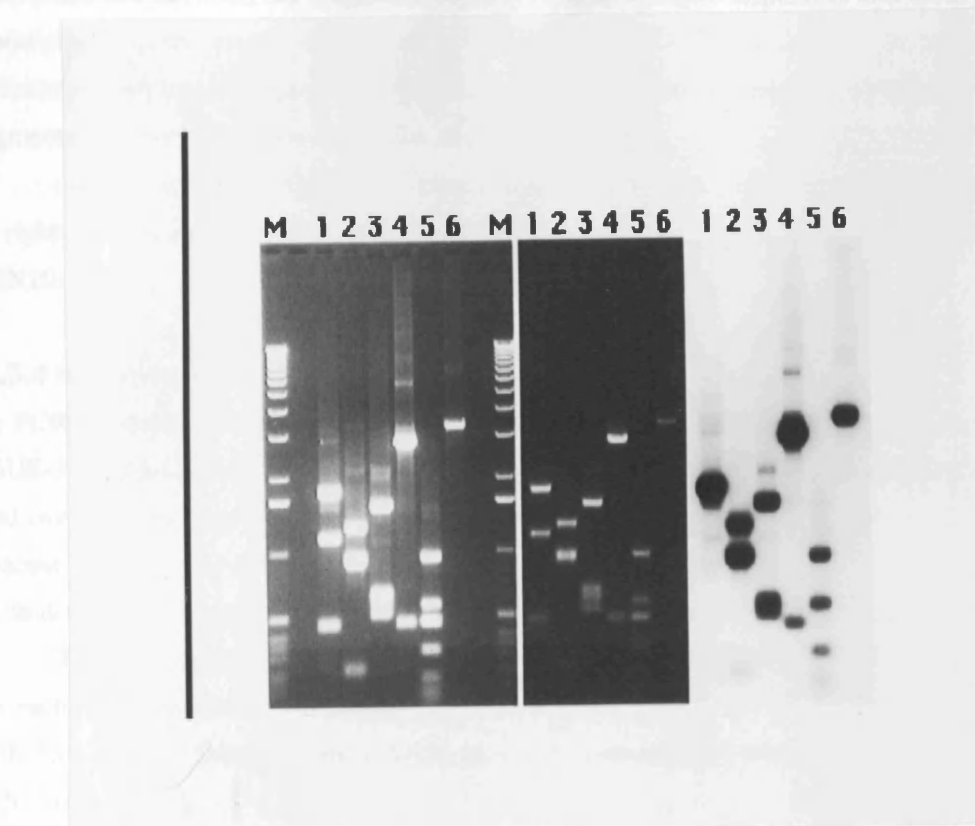
fragments

fragments

fragments

fragments

fragments



2.8.3). An autoradiogram of the blot (Fig. 6.13) confirmed that there was complete correspondence between the fragments expected to harbour *gus* sequences and those hybridizing with the probe. In the case of the two *EcoR* V fragments of 0.9 kb, which cannot be distinguished by their size, the hybridization patterns were deduced from those of the other fragments and the restriction map (Fig. 6.13).

As for the left junction, no evidence was found of a structurally altered *gus* gene at the right T-DNA junction. Two structurally intact *gus* genes might therefore be present in the AtEN101 plants.

6.3.3.4 Sequences of the junction segments

The PCR products corresponding to the right junction were partially sequenced with primers prGUS-3' and prGUS-4247 (TGAACAACGAACTGAACTGG). The sequences obtained with these two primers were essentially identical to the known sequence of *gus* although with some apparent base substitutions and deletions. These are likely to be due to PCR introduced errors and faults in the sequencing procedure rather than to actual mutations in the gene.

The left junction was also subjected to sequencing with primer LB-267 to determine its structure. It was found to be identical to the sequences in 5' RACE clones pSFR-4 and pSFR-15 with left border sequence being intact throughout, only truncated from position -19. At the start of the left border there was a five residues-long overlap with the conserved native plant genome sequence (see Section 6.2.4).

6.4 Summary

In this chapter the possible relation between the expression of fusion transcripts and the promoter of *PKT2* in AtEN101 was addressed. The investigative approach consisted of assessing the possible co-localization of the promoter(s) of the fusion transcripts and that of *PKT2*. Fusion transcripts were subjected to 5' RACE to obtain information on the region where they originated (Section 6.2). Several clones were obtained and sequenced. Analysis of the sequences revealed transcripts consisting of exon 1 of *PKT1* fused to sequences derived from a downstream LB region. This LB region had the opposite orientation to the LB region found upstream from *PKT2* in AtEN101. In this transgenic line the T-DNA had a complex structure, both its termini consisting of a LB region in a tail-to-tail orientation.

The complex T-DNA structure suggested that more than one *gus* reporter gene might be present and, therefore, potentially active in AtEN101. Southern blot analysis confirmed that two reporter genes were present in the line (Section 6.3.2). The required intactness of the reporters for detection of GUS activity was investigated by detailed restriction enzyme analysis of both the left junction (exon 1/*PKT1* region-upstream LB) and the right junction

(downstream LB-exon 1/*PKT2* region), as described in Section 6.3.3. For the analysis, the junction segments including nearly the entire *gus* genes were amplified by PCR. The results obtained showed that a reporter gene was located in each junction region, as expected.

Furthermore, they were compatible with the *gus* gene at the left junction being intact.

Similarly, restriction enzyme analysis and partial sequencing indicated that the reporter gene located at the right junction was also intact except, perhaps, at its 3' end.

Chapter 7

Identity of the *PKT1* and *PKT2* genes

7.1 Introduction

This chapter is concerned with the identification of the proteins encoded by the two wildtype transcripts mP18 and mP19, as well as the study of their functions through the investigation of their structural characteristics.

As described below (Section 7.2), each transcript was found to encode one long ORF designated P18-1 (mP18) and P19-1 (mP19) that were essentially identical. They were distinguished by the presence in P19-1 of an N-terminal extension of 43 residues relative to P18-1. Comparative analysis of these ORF sequences revealed that they were significantly similar to thiolase enzymes (Section 7.3), particularly to peroxisomal 3-ketoacyl-CoA thiolases (Section 7.4). The analysis also showed the presence in P18-1 and P19-1 of several previously defined primary structure features conserved in thiolases and identified others, as described in Section 7.5. To investigate further the probable function of the two proteins predictions of secondary structure were performed for P19-1 and compared with its known equivalent for the *S. cerevisiae* peroxisomal 3-ketoacyl-CoA thiolase (Section 7.6). Limited overall coincidence of the two was observed. Section 7.7 reports the study of the correspondence of P19-1 (and P18-1) to thiolases by homology-threading prediction of its 3D structure based on the known structure of the *S. cerevisiae* peroxisomal enzyme. Together the data indicated that P18-1 and P19-1 were peroxisomal 3-ketoacyl-CoA thiolases.

The likely cellular localization of P19-1 was congruent with the presence at the expected position in its N-terminal region of a match to the consensus for the peroxisomal targeting signal PTS-2 (Section 7.8). However, P18-1 lacked any obvious peroxisomal targeting signal. For future investigation of the cellular distributions of the two proteins, particularly of P18-1, plasmids encoding tagged versions of these proteins were designed and constructed (Section 7.8.3).

The comparative study of P18-1, P19-1 and the other thiolases permitted the development of diagnostic sequence elements (signatures) for the identification of these enzymes, as described in Section 7.9. This followed the recognition in thiolases of three blocks of sequence with highly conserved physico-chemical characteristics. One of the blocks was found to be superior to all three known thiolase signatures in the PROSITE database (Bairoch & Bucher, 1994) in its comprehensiveness of detection of the enzymes.

7.2 Identification of Open Reading Frames in the cDNA sequences

To assess the protein coding potential of mP18 and mP19 their sequences were conceptually translated in the three forward reading frames. The ORFs initiated with a methionine-codon were identified and some of these were selected for further analysis. The choice of ORFs was based on their length and relative positioning. In the initial stage, all ORFs longer than 100 codons were selected. A total of two ORFs satisfying this criterion were found for each of mP18 and mP19 (Table 7.1). The possibility that small ORFs (with potential regulatory functions) might be located upstream from these longer ORFs was then examined. Only one such ORF was found in mP19, P19-3, but it overlapped the longer P19-1 and was located far upstream (0.3 kb) from P19-2 (Table 7.1).

Upon examination, the longest ORF in mP18 (P18-1) was found to correspond to an N-terminal truncation of the P19-1 of mP19 (Fig. 7.1). The ORFs P18-2 and P19-2 were identical. These observations were not surprising given the close structural relationship between the two transcripts.

mP18				mP19			
ORFs	Frame	Length	Positions	ORFs	Frame	Length	Positions
P18-1	1	414 aa	94-1336	P19-1	2	457 aa	184-1555
P18-2	2	131 aa	401-794	P19-2	3	131 aa	620-1013
-	-	-	-	P19-3	3	83 aa	38-287

Table 7.1 Selected Methionine-initiated ORFs of the mP18 and mP19.

The characteristics of the ORFs chosen for further analysis are listed. The ORFs were detected by conceptual translation of the mP18 and mP19 sequences.

To establish which of the selected ORFs were more likely to encode proteins they were examined for a variety of structural features. The considerably longer size (>400 codons) of P18-1 and P19-1, and their positioning and extension relative to the length of the transcribed sequences mP18 (85.6%) and mP19 (81.1%), respectively, strongly indicated that they genuinely encoded proteins, much more so than the other, far smaller ORFs. The codon compositions of P19-1 and P18-2/P19-2 were determined and compared with a published codon usage table for *A. thaliana* (Nakamura *et al.*, 1997). A good correspondence was observed between the codon frequencies in P19-1 and the *Arabidopsis* codon usage (Fig. 7.2). By contrast, the P18-2/P19-2 sequence showed several incongruences relative to the expected

Figure 7.1 Alignment of the translated ORFs P18-1 and P19-1 with the sequence of mP19. The heptanucleotides that contain the first three methionine residues in P19-1 are shown in bold typeface. Underlined is the first methionine of the P18-1 translation (the third in P19-1). Residues absent at the N-terminus of P18-1 are shown in italics. The stop codon at the end of the ORFs is highlighted.

TGAGAACGATCGATCAAAAATCTCTCCCACGATAAAAATGAGG	43
ACCTGAAGCTCATTGTACTTAACGTCTTCCCCTTTCCAGAAAAATCAACCAAACCCATTTTCAGTTTCCC	113
CGGAAAGTTTTTGTCTTTCAATCACCATTAGAATTCAGCTGCAACACAAGTGTTCGCAACAGAGAGAATC	183
ATGGAGAGAGCTATGGAAAGACAAAAGATATTGCTTCGTATCTCAATCCAGTTTCTTCTTCTAATTCTT	253
M E R A M E R Q K I L L R H L N P V S S S N S S	24
CTCTTAAACATGAACCTTCTCTTCTGTCTCCTGTGAATTGTGTTTCTGAAGTTTCCCCCAATGGCTGCTTT	323
L K H E P S L L S P V N C V S E V S P M A A F	47
TGGAGATGACATTGTGATTGTAGCGGCATATCGTACCGCCATTTGCAAAGCGAGACGTGGAGGTTTCAAA	393
G D D I V I V A A Y R T A I C K A R R G G F K	70
GACACTCTTCTGATGATCTTCTTGCTTCTGTTCTTAAGGCTGTAGTGGAAGAACATCTTTGGATCCAA	463
D T L P D D L L A S V L K A V V E R T S L D P S	94
GTGAAGTTGGTGATATCGTTGTTGGTACCGTGATAGCGCCTGGTTCTCAGAGAGCAATGGAGGTAGAGT	533
E V G D I V V G T V I A P G S Q R A M E C R V	117
TGCAGCTTATTTTGTGGTTTTCTGACTCCGTGCCAGTTAGAACTGTCAATAGACAATGCTCATCAGGA	603
A A Y F A G F P D S V P V R T V N R Q C S S G	140
CTACAAGCAGTTGCTGATGTTGCTGCTTCCATTAGAGCTGGTTATTACGACATTGGTATTGGTGCTGGAG	673
L Q A V A D V A A S I R A G Y Y D I G I G A G V	164
TGGAATCAATGTCAACTGATCATATTCTGGAGGCGGCTTTCATGGCTCTAATCCAAGAGCACAGGATTT	743
E S M S T D H I P G G G F H G S N P R A Q D F	187
CCCAAAAGCCCGTGATTGTTTGCTTCCAATGGGAATTACTTCTGAAAACGTTGCAGAAAGGTTCCGGTGTC	813
P K A R D C L L P M G I T S E N V A E R F G V	210
ACAAGAGAAGAGCAAGATATGGCTGCGGTGGAGTCTCACAAACGCGCTGCAGCTGCAATCGCGTCTGGTA	883
T R E E Q D M A A V E S H K R A A A A I A S G K	234
AACTCAAGGATGAAATCATTCCTGTGTCTACTAAGATTGTGGACCCTGAGACTAAAGCAGAGAAGGCAAT	953
L K D E I I P V A T K I V D P E T K A E K A I	257
CGTCGTATCTGTTGATGACGGTGTACGTCCAAACTCAAACATGGCAGATTTGGCAAAGCTGAAGACTGTC	1023
V V S V D D G V R P N S N M A D L A K L K T V	280
TTTAAACAGAACGTTCCACCACAGCTGGCAATGCTAGTCAGATCAGTGATGGTGCTGGAGCAGTACTGC	1093
F K Q N G S T T A G N A S Q I S D G A G A V L L	304
TAATGAAGAGAAGTTTGGCTATGAAGAAGGGACTTCCCATTTCTGGAGTATTCAGGAGTTTTGCTGTTAC	1163
M K R S L A M K K G L P I L G V F R S F A V T	327
TGGTGTGGAACCATCTGTAATGGGTATTGGTCCAGCTGTTGCCATTCCCGCTGCAACTAAGCTCGCAGGG	1233
G V E P S V M G I G P A V A I P A A T K L A G	350
CTCAACGTCAGCGATATTGATCTATTTCGAGATCAATGAGGCATTTGCATCTCAGTATGTGTACTCTTGCA	1303
L N V S D I D L F E I N E A F A S Q Y V Y S C K	374
AGAAGTTAGAGCTGGATATGGAAAAGGTCAATGTTAATGGAGGAGCCATTGCTATTGGCCATCCCCTGGG	1373
K L E L D M E K V N V N G G A I A I G H P L G	397
TGCTACAGGAGCTCGATGTGTTGCGACATTGTTGCACGAGATGAAGCGGAGAGGAAAAGATTGCCGCTTT	1443
A T G A R C V A T L L H E M K R R G K D C R F	420
GGAGTAATCTCAATGTGCATAGGCACTGGAATGGGAGCTGCAGCTGTTTTTGAGAGGGGAGACTCTGTTG	1513
G V I S M C I G T G M G A A V F E R G D S V D	444
ATAACTTGTCCAACGCTCGTGTGGCTAACGGGGATAGTCATTAGAACATCGAAGAGAGCTTGAATAAGTA	1583
N L S N A R V A N G D S H <	457
GAAGTAATGATGCATTGAGTCTAATAAATATGATGCTTTAGCTCTTTCACATTGCTGAACAATGAAAAC	1653
TTTGTCAATTCTGAGTTTAAAATCAACTACTTTTCTCTG	1691

codon distribution for a genuine gene (Fig. 7.2). Among others, it was noted that of the 8 serine codons present none was UCU although these account for 39.9% of serine codons in *Arabidopsis* genes. P19-2 had a single codon for glycine and no aspartate codons were present. The average frequencies of codons for these two aminoacids are 6.9% and 5.4% respectively. Also 14 CUG codons, of a total of 22 encoding leucine, were counted in P19-2. Yet, in *Arabidopsis*, CUG codons only represent 16.4% of the total leucine codons. In addition to codon frequencies, the sequences surrounding the putative translation initiation sites were examined. From a compilation of 957 sequences containing known or putative translation initiation sites of *A. thaliana* genes (Brown *et al.*, 1993; Dalphin *et al.*, 1998) the heptanucleotides surrounding the initiation codon were extracted and their frequencies determined. It was observed that the heptanucleotides CAATGTG and CGATGTG containing the first Met-codon of P18-2/P19-2 and P19-3, respectively, do not occur in the list of translation initiation sites of *Arabidopsis*. The absence has a relatively low probability of occurrence of 0.0236 in a sample of 957 sites (Table 7.2). This is particularly significant

Met-codon	Heptanucleotide	Occurrences	Probability
1st Met of P18-1	CAATGGC	60	4.166×10^{-49}
1st Met of P19-1	TCATGGA	8	0.0224
2nd Met of P19-1	CTATGGA	2	0.1662
1st Met of P18-2/P19-2	CAATGTG	0	0.0236
1st Met of P19-3	CGATGTG	0	0.0236
-	-	3	0.2074
-	-	4	0.1940

Table 7.2 Probabilities of the observed frequencies of heptanucleotides containing the translation initiation codon.

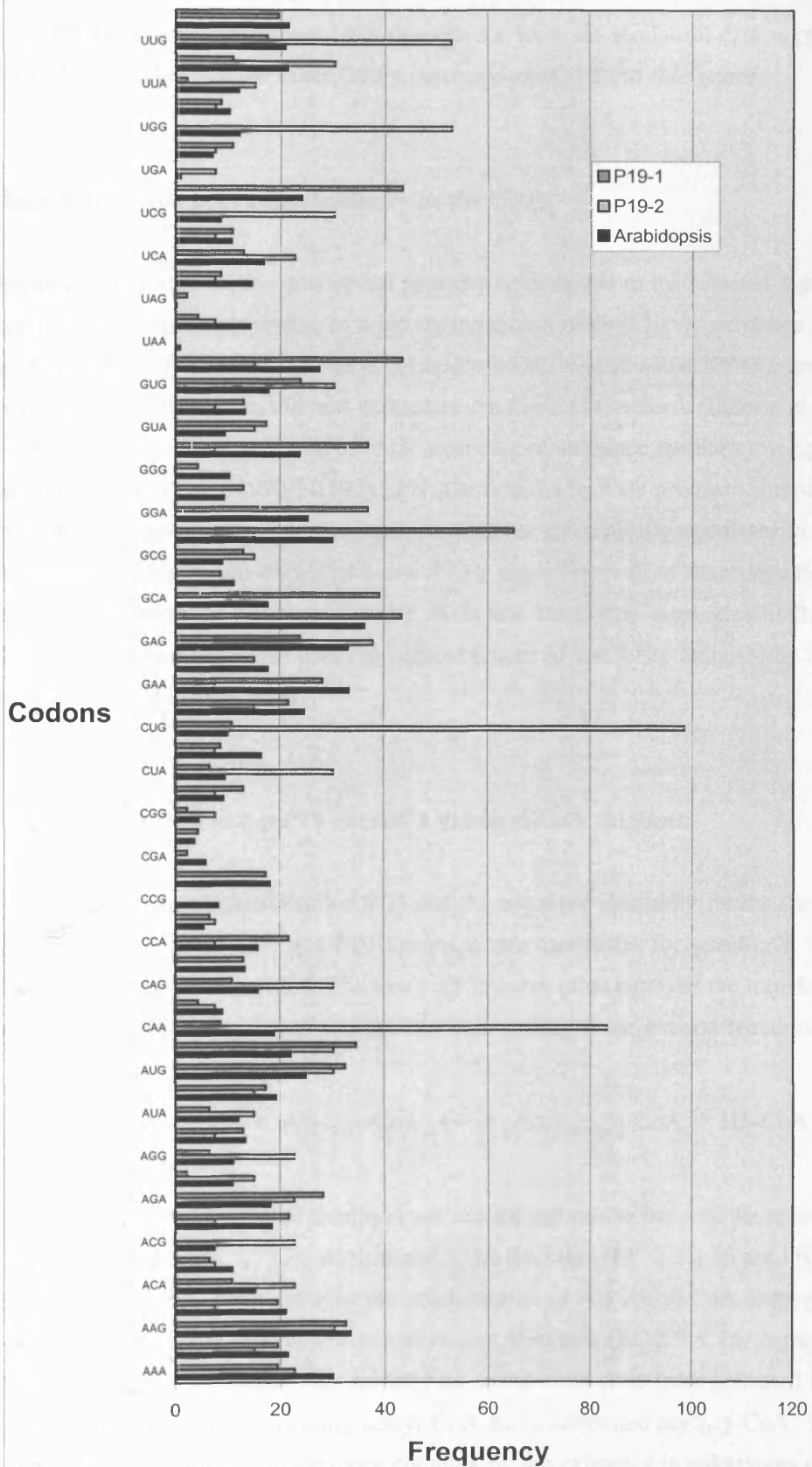
The calculated probabilities refer to the frequencies computed for the present work from a list of 957 sequences containing the translation initiation sites of *A. thaliana* genes (Brown *et al.*, 1993; Dalphin *et al.*, 1998). The source of the heptanucleotides in the ORFs of mP18 and mP19 and their sequences are indicated. The number of occurrences in the list is shown together with its associated probability. The expected average frequency of any given heptanucleotide would be 3.734. The probabilities for the two closest consecutive integers containing this value, 3 and 4, are shown for comparison.

in view of the observed bias in the composition of such sites (data not shown). Indeed, the heptanucleotide CAATGGC containing the first Met-codon of P18-1 (the third in P19-1; cf. Fig. 7.1) has 60 occurrences in the same compilation. The equivalent heptanucleotide in P19-

Figure 7.2 Comparative diagram of codon usage for the ORFs P19-1 and P19-2.

A bar diagram of codon frequencies in the ORF P19-1 and P19-2 and the average frequencies in *A. thaliana* is shown. The averages frequencies for *A. thaliana* correspond to a total of 3533 protein coding sequences representing 1497366 codons (Nakamura *et al.*, 1997).

Comparative Codon Usage



1, TCATGGA occurs 8 times in the extracted list. Four codons downstream, the next Met-codon of P19-1 is embedded in CTATGGA, a sequence that was only found twice at translation initiation sites in *Arabidopsis*.

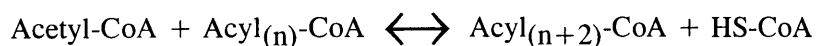
In summary, the various lines of evidence from the structural data suggested that P18-1 and P19-1, but not the other ORFs, corresponded to *bona fide* genes.

7.3 Search for sequences with similarity to the ORFs

A significant similarity of the conceptual peptides represented in the selected segments to known or putative proteins would be a strong indication of their likely existence and suggestive of their function. Therefore, all selected ORFs were subjected to comparison with the sequences represented in the non-redundant database of GenBank (Benson *et al.*, 1997). The search was based on detection of local segments of sequence similarity using a BLAST program (Altschul *et al.*, 1990), TBLASTN (Section 2.11). This program compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames. Of the select group only P18-1 and P19-1 were observed to share significant similarities to known and putative proteins. With few exceptions these identified proteins were thiolase enzymes, including all with the highest scores of similarity assigned by the TBLASTN algorithm (Fig. 7.3).

7.4 Transcripts mP18 and mP19 encode 3-ketoacyl-CoA thiolases

Together, the structural data (Section 7.2) and the sequence similarity results (Section 7.3) indicated that the ORFs P18-1 and P19-1 were strong candidates for genuine *A. thaliana* genes encoding thiolase enzymes. Thiolases are enzymes that catalyze the transfer of an acetyl moiety between coenzyme A and an acyl-CoA, according to the general reaction:



Two distinct groups of thiolases are recognized on the basis of the specific reactions they catalyze (Middleton, 1973). Acetoacetyl-CoA thiolases (EC 2.3.1.9) are involved in biosynthetic pathways. They catalyze the condensation of two acetyl-CoA units generating acetoacetyl-CoA. On the contrary, 3-ketoacyl-CoA thiolases (EC 2.3.1.16) participate in the catabolism of fatty acids where they transfer an acetyl equivalent from activated fatty acids (acyl_n-CoA) to coenzyme A forming acetyl-CoA and a shortened acyl_{n-2}-CoA. This simple biochemical classification is made more complex by the existence in eukaryotes of isozymes

TBLASTN of P19-1

Query = P19-1 (457 letters)

Database: Non-redundant GenBank+EMBL+DDBJ+PDB sequences

Sequences producing High-scoring Segment Pairs:		Reading Frame	High Score	Smallest Sum Probability P(N)
dbj D70895 D70895	Cucurbita sp. mRNA for 3-ketoac...	+1	1094	1.7e-214 2
emb X93015 BNGBHTPRE	B.napus mRNA for glyoxysomal be...	+1	1071	2.3e-142 2
emb X67696 CS3KCT	C.sativus mRNA for 3-ketoacyl-C...	+2	1101	5.6e-142 1
emb X75329 MITHMF5	M.indica (Manila) THMF5 mRNA fo...	+1	1015	1.6e-132 2
emb X14813 HS3OCT	Human liver mRNA for 3-oxoacyl-...	+1	538	5.6e-132 5
emb X12966 HS3OCOAT	Human mRNA for 3-oxoacyl-CoA pe...	+1	538	5.6e-132 5
gb M32801 RATKCOAT	Rat 3-ketoacyl-CoA thiolase 2 m...	+1	535	6.1e-130 5
gb J02749 RATKPCOAT	Rat peroxisomal 3-ketoacyl-CoA...	+2	537	4.0e-129 5
gb AF035295 AF035295	Homo sapiens clone 23623 mRNA, ...	+1	428	2.2e-115 6
dbj D17321 D17321	C.tropicalis gene for 3-ketoacy...	+1	400	1.9e-98 5
dbj D17320 D17320	C.tropicalis gene for 3-ketoacy...	+1	400	4.8e-98 5
emb X69988 YLTHI	Y.lipolytica gene for 3-oxoacyl...	+2	474	1.1e-88 3
emb Z47047 SCCHRIX	S. cerevisiae chromosome IX com...	-3	400	1.5e-87 5
emb Z38059 SC4357	S.cerevisiae chromosome IX lamb...	-2	400	1.6e-87 5
emb X53395 SCPOT1	S.cerevisiae POT1 gene for 3-ox...	+2	400	1.7e-87 5
emb X53946 SCFOX3	S.cerevisiae gene for peroxisom...	+2	400	1.7e-87 5
dbj D90851 D90851	E.coli genomic DNA, Kohara clon...	+2	397	1.9e-71 5
gb AE000311 ECAE000311	Escherichia coli K-12 MG1655 se...	+2	397	1.9e-71 5
emb Z92974 TTBCSOPRN	T.thermosaccharolyticum BCS ope...	+2	408	1.5e-70 3
emb Z82038 CTZ82038	C.thermosaccharolyticum etfB, e...	+2	408	1.5e-70 3
gb AE001021 AE001021	Archaeoglobus fulgidus section ...	-3	392	5.3e-70 3
gb U08465 CAU08465	Clostridium acetobutylicum ATCC...	+1	365	1.2e-69 4
gb L37761 ACCPHAABC	Acinetobacter sp. (strain RA384...	+3	408	1.9e-69 3
gb U37723 FNU37723	Fusobacterium nucleatum immunos...	+1	200	4.1e-68 6
gb L01113 TSSPHBAC	Thiocystis violacea beta-ketoth...	-1	355	7.0e-68 5
gb S54369 S54369	phbCAB operon: phbA=beta-ketoth...	-1	355	7.0e-68 5
dbj D16294 HUMDSAEC	Human mRNA for mitochondrial 3-...	+1	404	1.9e-66 3
gb U32761 HIU32761	Haemophilus influenzae from bas...	-3	409	3.5e-66 3
dbj D49362 PDEPHAA	Paracoccus denitrificans genes ...	+2	346	4.0e-66 5
gb J04987 AFAKTLAACA	A.eutrophus beta-ketothiolase (...)	+1	355	4.1e-66 4
gb AF002013 AF002013	Alcaligenes sp. SH-69 beta-keto...	+2	374	4.1e-66 4
gb AE000582 HPAE000582	Helicobacter pylori section 60 ...	+1	388	5.3e-66 5
gb L01112 CVNPHB	Chromatium vinosum poly(3-hydro...	+1	354	2.5e-65 5
emb X05341 RN3OCTR	Rat mRNA for 3-oxoacyl-CoA thio...	+3	381	2.5e-63 3
gb J02631 ZOGTHIZR	Z.ramigera thiolase gene, compl...	+3	374	7.3e-63 4
gb U17226 RMU17226	Rhizobium meliloti beta-ketothi...	+3	361	1.5e-61 5

Figure 7.3 Output of BLAST comparison of P19-1 to the non-redundant GenBank database.

The sequence of the P19-1 ORF was compared on-line with those deposited at the database (on July 1997). The algorithm employed was TBLASTN (Altschul *et al.*, 1990) using the similarity matrix BLOSUM62 (Henikoff & Henikoff, 1992). Only the highest scoring hits are shown. All correspond to sequences encoding known or putative thiolase enzymes.

of both types which are active in different cellular compartments. Originally described in the cytosol and mitochondria Middleton (1973) they were later found to be present also in peroxisomes (Lazarow, 1978).

To obtain indications on the specific function of the putative P18-1 and P19-1 proteins each was individually aligned with sequences to which the BLAST search uncovered strong similarities, as well as others recently described. For these pairwise alignments the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) was utilized, as implemented by the GAP procedure of the GCG package (Devereux *et al.*, 1984; Genetics Computer Group, 1994). This algorithm finds the optimal global alignment between two sequences for a given set of parameters. A subset of the results attained for the P19-1 pairwise alignments is summarized in Table 7.3. Noteworthy are the comparable lengths of P18-1 and P19-1 to those of the thiolases, about 400-500 residues. Furthermore, the similarity between the two putative *Arabidopsis* proteins to thiolases was not limited to a domain but rather extended throughout their length, as is exemplified in Figure 7.4 for P19-1 and the *S. cerevisiae* peroxisomal 3-ketoacyl-CoA thiolase.

It was clear from the analysis that amongst the aligned proteins the highest pairwise similarity occurred between P19-1 and peroxisomal 3-ketoacyl-CoA thiolases. This is exemplified by the results listed in Table 7.3 obtained with all known plant thiolases as well as of organisms representative of mammals, yeasts and bacteria. The identity levels of P19-1 to the peroxisomal EC 2.3.1.16 enzymes ranged from 45% to 72% (for a gap penalty of 3.0 and gap extension penalty of 0.1) but were less than 40% to their mitochondrial equivalents, and only 34-44% relative to acetoacetyl-CoA thiolases. A better illustration is achieved by considering the pairwise comparisons between P19-1 and the various thiolase types from a single organismal source, using a same set of parameters. For example, relative to the human enzymes, P19-1 was found to be 55.9% identical and 70.5% similar to the peroxisomal 3-ketoacyl-CoA thiolase (gap penalty = 3.0; gap extension penalty = 0.1). These values were reduced to 42.7% and 59.1%, respectively, in relation to the human cytosolic acetoacetyl-CoA thiolase. The human mitochondrial EC 2.3.1.16 appeared to be slightly more dissimilar with 38.1% of identical and 58.5% similar residues. The most dissimilar of the human enzymes listed was the mitochondrial acetoacetyl-CoA thiolase which shared with P19-1 35.2% identical and 57.8% similar residues. Even lower values were obtained relative to a second isozyme of the mitochondrial 3-ketoacyl-CoA thiolase (Kamijo *et al.*, 1994) and to the sterol carrier protein X (Ohba *et al.*, 1994) with only 33.5% and 26.0% identity, respectively (data not shown).

The levels of similarity and identity obtained with the GAP procedure are dependent on the particular set of parameters used to generate the alignments. Also, different pairwise alignments created by GAP are not strictly comparable. Nevertheless, the trend revealed in Table 7.3 is relevant as changing the GAP algorithm parameters did not but change the results

GAP alignment

P19-1 vs. *S. cerevisiae* 3-ketoacyl-CoA thiolase

Gap Weight: 3.000 Length Weight: 0.100
Similarity: 61.858% Identity: 45.232%

```

P19-1      1  MERAMERQKILLRHLNPVSSNSSLKHE--PSLLSPVNCVSEVSPMAAFG 48
              :| . |:||.. .| ::::.. .| :. .
S.c.       1  MSQRLQSIKDHLVESAMGKGESKRKNSLLEKRP 33

P19-1     49  DDIVIVAAYRTAICKARRGGFKDTLPDDLASVLKAVVERTS----LDPS 94
              :|:||||| |.|||.|: :|:||||. .| || ..|...:| . | .
S.c.      34  EDVVIVAANRSAIGKGFKAQKDVNTDYLLYNFLNEFIGRFPEPLRADLN 83

P19-1     95  EVGDIVVGTVIAPGSQRAMECRVAAYFAGFPDVPVVRTVNRQCSSGLQAV 144
              ::::..|.:|..|.. | | |.|. .|:| |.|. .:||||||| ||
S.c.      84  LIEEVACGNVLNVGAG-ATEHRAACLASGIPYSTPFVALNRQCSSGLTAV 132

P19-1    145  ADVAASIRAGYYDIGIGAGVESMSTDHIPGGGFH-GSNPRAQDFPKARDC 193
              .|:|..|.:| .|||:| |||||... ..:| .. |. .:|:|
S.c.     133  NDIANKIKVGQIDIGLALGVESMTNNYKNVNPLGMISSEELQKNREAKKC 182

P19-1    194  LLPMGITSENVAFERFGVTREEQDMAAVESHKRAAAAIASGKLKDEIIPVA 243
              |:|||||.||||..| :|:|:| |.:.|:| | ..| :.||||:|
S.c.     183  LIPMGITNENVAANFKISRKDQDEFANASYQKAYKAKNEGLFEDEILPI- 231

P19-1    244  TKIVDPETKAEKAIVVSVDDGVRPNSNMADLAKLKTVF-KQNGSTTAGNA 292
              | ..: :. |:|.||| . ..|...:..| |:|.|||||
S.c.     232  -----KLPDGSICQSDEGPRPNVTAESLSSIRPAFIKDRGTTTAGNA 273

P19-1    293  SQISDGAGAVLLMKRSLAMKKGLPILGVFRSFVAVTGVEPSVMGIGPAVAI 342
              ||:||||.:||| :||:| . .||:| | : .|...||.|.:||:| ||
S.c.     274  SQVSDGVAGVLLARRSVANQLNLPVLGRYIDFQTVGVPEIMGVGPAYAI 323

P19-1    343  PAATKLAGLNVSDIDLFEINEAFASQYVYSCKKLELDMEKVN NVNGGAIAI 392
              | . . .||. | |||:|||||||. | :|:..|:|:|:|:|..| ||||:
S.c.     324  PKVLEATGLQVQDIDIFEINEAFAAQALYCIHKLGLIDLNVNPRGGAIAL 373

P19-1    393  GHPLGATGARCVATLLHEMKRRGKDCRFVISMCI GTGMGAAAVFERGDS 442
              |||||.|||| |||:|:|:|: .:|:|:||||||| ||||:| :
S.c.     374  GHPLGCTGARQVATILRELKKD----QIGVSMCI GTGMGAAAFI KE 417

```

Figure 7.4 Alignment of the ORF P19-1 with the *S. cerevisiae* peroxisomal 3-ketoacyl-CoA thiolase.

The alignment was produced with GAP (Devereux *et al.*, 1984; Genetics Computer Group, 1994). A gap penalty of 3.0 and a gap extension penalty of 0.1 were applied. The comparison matrix was the normalized table of Dayhoff (Gribskov & Burgess, 1986). Solid bars link identical residues in the two sequences. Single and double dots indicate similar residues. Gaps are shown as dashes. The *S. cerevisiae* sequence is labelled *S.c.* The N-terminal 43 residues absent from P18-1 are in italics.

marginally (Table 7.3; data not shown). For example, an increment of the gap penalty from 3.0 to 10.0 (resulting in a decrease of the number of gaps) caused variations of only between 0.0-6.4% in the levels of identical and similar residues (Table 7.3). Most importantly, the qualitative ranking of similarity of P19-1 to the various thiolases was not altered by parameter changes. Also, the relative quality of the alignments as measured by several criteria was found to be superior for those associated with higher levels of identical and similar residues (*e.g.* Table 7.3; data not shown). To exemplify, the metric (so-called Quality) of the GAP alignments between P19-1 and the *H. sapiens* thiolases varied from 277.6 (mitochondrial EC 2.3.1.16) to 406.3 (peroxisomal EC 2.3.1.16) for a gap penalty of 3.0 (Table 7.3). However, the metrics of similar alignments of P19-1 to the same sequences randomly shuffled only varied from about 137.8 ± 5.6 (mitochondrial EC 2.3.1.16) to 145.0 ± 3.7 (peroxisomal EC 2.3.1.16).

That P19-1 showed a higher degree of similarity to peroxisomal 3-ketoacyl-CoA thiolases, some of which had their activity experimentally demonstrated, strongly indicated that they are homologous and indeed that they have similar functions. This is particularly true for the *Brassica napus* peroxisomal thiolase (70% identical residues; Olesen *et al.*, 1997) and for PKT3 of *A. thaliana* (Newman *et al.*, 1994), which was recently proven to be a peroxisomal enzyme (Hayashi *et al.*, 1998). Nevertheless, shorter than testing the activities of the enzymes themselves, only a more detailed analysis to determine the conservation of functionally important residues and secondary and tertiary structure features can give additional insights into the role of the protein.

7.5 Conserved primary structure features in P18-1 and P19-1

Thiolases catalyze Claisen condensations/cleavages through a ping-pong mechanism involving the formation of an enzyme thioester intermediate (Thompson *et al.*, 1989). Thompson *et al.* (1989) were the first to map the formation of the enzyme thioester intermediate to a particular cysteine residue (Cys89) in the acetoacetyl-CoA thiolase of *Zoogloea ramigera*. Another cysteine residue (Cys378) is also required for the cleavage of the intermediate (Palmer *et al.*, 1991; Williams *et al.*, 1992). An additional residue, His347, contributes to catalysis by the *Z. ramigera* enzyme (Masamune *et al.*, 1989; Williams *et al.*, 1992). The two cysteines and surrounding sequences are well conserved in thiolases (Igual *et al.*, 1992). To ascertain if these cysteines are also conserved in P19-1 (and P18-1), as well as to identify conserved and non-conserved segments in these proteins, a multiple alignment with thiolase sequences was performed. The sequences included those detected by the BLAST program (Section 7.3) and others retrieved from the GenBank database. Represented were thiolases of all types and of eukaryotic organellar origin derived from plants, animals and yeasts, as well as from

Species/ Type	<i>B. napus</i> perox. EC 2.3.1.16		<i>Cucumis sativus</i> perox. EC 2.3.1.16		<i>Cucurbita sp.</i> perox. EC 2.3.1.16	
Gap Pen.	3.0	10.0	3.0	10.0	3.0	10.0
Ident.	70.5 %	69.2 %	72.3 %	70.6 %	71.7 %	70.1 %
Simil.	84.7 %	82.6 %	85.0 %	83.4 %	84.3 %	82.9 %
Gaps	4	2	6	3	3	1
Quality	530.8	516.2	528.5	506.2	536.3	527.4
Av.Qual.	151.5 ± 3.7	116.8 ± 6.5	152.5 ± 5.5	113.4 ± 4.5	150.2 ± 4.0	113.9 ± 3.0
Length	462 aa		459 aa		461 aa	
Refs.	(Olesen <i>et al.</i> , 1997)		(Preisig-Müller & Kindl, 1993)		(Kato <i>et al.</i> , 1996)	

Species/ Type	<i>Mangifera indica</i> perox. EC 2.3.1.16		<i>A. thaliana</i> perox. EC 2.3.1.16 (PKT3)		<i>Raphanus sativus</i> cytos. EC 2.3.1.9	
Gap Pen.	3.0	10.0	3.0	10.0	3.0	10.0
Ident.	71.4 %	68.0 %	71.4 %	69.4 %	38.2 %	33.2 %
Simil.	84.5 %	81.3 %	86.0 %	83.3 %	59.7 %	53.3 %
Gaps	4	1	5	2	12	5
Quality	506.6	473	533.6	517.4	258.1	204.4
Av.Qual.	144.6 ± 5.4	103.7 ± 4.2	150.6 ± 4.3	113.6 ± 3.8	141.6 ± 3.5	107.4 ± 4.0
Length	430 aa		462 aa		406 aa	
Refs.	(Bojorquez & Gomez-Lim, 1995)		(Hayashi <i>et al.</i> , 1998)		(Vollack & Bach, 1996)	

Species/ Type	<i>Rattus norvegicus</i> perox. EC 2.3.1.16 (A)		<i>R. norvegicus</i> mitoch. EC 2.3.1.16		<i>R. norvegicus</i> mitoch. EC 2.3.1.9	
Gap Pen.	3.0	10.0	3.0	10.0	3.0	10.0
Ident.	56.8 %	56.3 %	37.2 %	35.6 %	34.4 %	34.4 %
Simil.	71.5 %	71.0 %	57.4 %	55.4 %	58.1 %	55.0 %
Gaps	6	5	6	4	10	6
Quality	411.8	375.9	275.2	245.9	271.4	226.5
Av.Qual.	148.6 ± 4.6	112.8 ± 3.4	136.6 ± 3.0	102.6 ± 6.1	145.5 ± 3.5	111.5 ± 3.4
Length	434 aa		397 aa		424 aa	
Refs.	(Hijikata <i>et al.</i> , 1990) (Bodnar & Rachubinski, 1990)		(Arakawa <i>et al.</i> , 1987)		(Fukao <i>et al.</i> , 1989)	

Species/ Type	<i>H. sapiens</i> perox. EC 2.3.1.16		<i>H. sapiens</i> mitoch. EC 2.3.1.16		<i>H. sapiens</i> mitoch. EC 2.3.1.9	
Gap Pen.	3.0	10.0	3.0	10.0	3.0	10.0
Ident.	55.9 %	55.8 %	38.1 %	36.1 %	35.2 %	31.0 %
Simil.	70.5 %	70.5 %	58.5 %	54.9 %	57.8 %	50.0 %
Gaps	5	5	6	3	12	4
Quality	406.3	371.3	277.6	249.8	268.4	221.6
Av.Qual.	145.0 ± 3.7	111.3 ± 5.3	137.8 ± 5.6	94.8 ± 2.9	145.0 ± 3.7	110.0 ± 4.1
Length	424 aa		397 aa		427 aa	
Refs.	(Bout <i>et al.</i> , 1988)		(Abe <i>et al.</i> , 1993)		(Fukao <i>et al.</i> , 1990)	

Species/ Type	<i>H. sapiens</i> cytos. EC 2.3.1.9		<i>S. cerevisiae</i> perox. EC 2.3.1.16		<i>S. cerevisiae</i> cytos. EC 2.3.1.9	
Gap Pen.	3.0	10.0	3.0	10.0	3.0	10.0
Ident.	42.7 %	40.3 %	45.2 %	43.4 %	34.7 %	30.6 %
Simil.	59.1 %	55.8 %	61.9 %	59.4 %	57.7 %	51.8 %
Gaps	8	5	7	5	11	4
Quality	287.5	244.2	323.9	281.8	250.0	208.4
Av.Qual.	143.0 ± 6.9	104.5 ± 2.5	137.4 ± 3.8	103.9 ± 4.4	138.6 ± 4.0	105.6 ± 3.8
Length	397 aa		417 aa		398 aa	
Refs.	(Song <i>et al.</i> , 1994)		(Igual <i>et al.</i> , 1991)		(Hiser <i>et al.</i> , 1994)	

Species/ Type	<i>Yarrowia lipolytica</i> perox. EC 2.3.1.16		<i>Candida tropicalis</i> perox. EC 2.3.1.16 (A)		<i>C. tropicalis</i> perox. EC 2.3.1.9	
Gap Pen.	3.0	10.0	3.0	10.0	3.0	10.0
Ident.	52.4 %	52.2 %	45.6 %	45.1 %	36.9 %	31.2 %
Simil.	66.2 %	64.7 %	63.5 %	62.7 %	56.4 %	59.7 %
Gaps	6	4	5	4	11	3
Quality	366.8	336.5	338.4	310.1	259.9	222.1
Av.Qual.	142.6 ± 4.0	111.7 ± 3.7	105.7 ± 2.9	105.7 ± 2.2	141.5 ± 4.5	106.3 ± 4.8
Length	414 aa		408 aa		403 aa	
Refs.	(Berninger <i>et al.</i> , 1993)		(Kanayama <i>et al.</i> , 1994)		(Kurihara <i>et al.</i> , 1992)	

Species/ Type	<i>E. coli</i> EC 2.3.1.16		<i>E. coli</i> EC 2.3.1.9		<i>Zoogloea ramigera</i> EC 2.3.1.9	
Gap Pen.	3.0	10.0	3.0	10.0	3.0	10.0
Ident.	43.3 %	42.1 %	44.1 %	43.5 %	41.7 %	40.9 %
Simil.	61.2 %	59.4 %	62.7 %	60.7 %	59.9%	57.3 %
Gaps	8	6	7	5	6	4
Quality	280.8	236.3	302.4	265.0	289.5	256.2
Av.Qual.	137.3 ± 4.7	101.2 ± 4.8	140.2 ± 3.0	107.7 ± 4.4	139.0 ± 3.8	104.5 ± 2.5
Length	387 aa		394 aa		391 aa	
Refs.	(DiRusso, 1990)		(Itoh <i>et al.</i> 1996)		(Peoples <i>et al.</i> , 1987)	

Table 7.3 Partial summary of alignments between P19-1 and known thiolase enzymes.

The alignments were performed with GAP (Section 2.11) with a gap penalty of 3.0 or 10.0, as indicated, and a gap extension penalty of 0.1. A normalized table of Dayhoff (Gribskov & Burgess, 1986) was used as comparison matrix. Shown are the percentage of identical (**Ident.**) and similar (**Simil.**) residues, as well as the number of gaps observed between each pair of sequences. Also indicated is the type of thiolase and, for eukaryotes, its cellular location in the cytosol (**cytos.**), mitochondria (**mitoch.**) or peroxisomes (**perox.**). For estimation of the quality of the alignments the value of the metric used by GAP to choose the optimal alignment is shown (**Quality**). Also shown for each comparison are the average metrics of 20 alignments of P19-1 to random shuffles of the target sequences (**Av. Quality**). The standard deviations associated with the Average Quality values are given.

eubacteria and archaeobacteria. For the alignment the PILEUP routine of the GCG suite of programs (Devereux *et al.*, 1984; Genetics Computer Group, 1994) was used.

Close inspection of the multiple alignment in Figure 7.5 showed that positions recognized as being conserved in the heterologous sequences (Igual, #17 1992), were similarly occupied by conserved residues in P19-1. In particular, both cysteine residues found to be catalytically necessary in thiolases (Masamune *et al.*, 1989; Palmer *et al.*, 1991; Williams *et al.*, 1992) are present in P19-1 (Cys₁₃₇ and Cys₄₂₆) and P18-1 (Cys₉₄ and Cys₃₈₃), each embedded in a conserved segment (Fig. 7.5). These two segments include, in P19-1, the sequences 132-151 and 421-434 that conform to two of the three signatures for thiolase enzymes recognized in the PROSITE database of conserved protein motifs (Bairoch & Bucher, 1994), as shown in Figure 7.6. Signatures are defined as minimal segments of sequence that are common to most or all members of a protein family and either are absent or rarely occur in other proteins. The other PROSITE thiolase signature (signature 2) is located in the C-terminal region, between residues 384-400 in the P19-1 sequence. It contains the conserved histidine residue that contributes to catalysis (Williams *et al.*, 1992), which in P18-1 and P19-1 corresponds to His₃₅₁ and His₃₉₄, respectively. Therefore, both *A. thaliana* proteins shared with known thiolases the functionally most important and some of the best preserved segments of these enzymes at the level of primary structure.

Examination of the multiple alignment revealed several other highly conserved blocks of sequence. Prominent amongst these were the blocks A, B, and C corresponding to the sequences DIGIGAGVESM, DGAGAVLL and DIDLFEINEAFA at positions 157, 297 and 355 of P19-1, respectively (Fig 7.5). Their potential use as diagnostic sequences for thiolases is further explored in Section 7.9.

7.6 Prediction of secondary structure features

The peroxisomal 3-ketoacyl-CoA thiolase of *S. cerevisiae* is the structurally best characterized enzyme of its type having being resolved crystallographically at 2.8 Å (Mathieu *et al.*, 1994) and 1.8 Å resolutions (Mathieu *et al.*, 1997). For this reason it was used as a paradigm to which to compare models of secondary structure for P18-1 and P19-1.

A variety of methods were employed to develop models of secondary structure for P19-1 (Section 2.11). Included were an improved GOR statistical method of Garnier *et al.* (Garnier *et al.*, 1996; Gibrat *et al.*, 1987), the homology method of Levin *et al.* (Levin *et al.*, 1986), the double prediction method (DPM) of Deleage and Roux (1987a; 1987b), the statistical method for non-local interactions PREDATOR (Frishman & Argos, 1996), a neural-network method for multiple-aligned sequences, SOPMA (Geourjon & Deleage, 1994; 1995), and the PHD neural-network method (Rost & Sander, 1993, 1994; Rost *et al.*, 1994). The results of these secondary structure analysis are summarized in Figure 7.7 where

Figure 7.5 Multiple alignment of P19-1 with known or putative thiolases sequences.

The alignment was performed with the PILEUP routine of the GCG package (Devereux *et al.*, 1984). The parameters used were: gap penalty=3.0; gap extension penalty=0.1. The comparison matrix was the normalized table of Dayhoff (Gibbskov & Burgess, 1986). Shaded are conserved positions (including those where conservative substitutions occur). The shading tone indicates the level of conservation: Black-100%; Dark grey- >80%; Light grey- >60%. The three PROSITE signatures (Bairoch & Bucher, 1994) and the conserved blocks A, B and C identified in this work are identified. Coloring was performed with GeneDoc (Nicholas *et al.*, 1997). The following abbreviation format was used to describe the sequences: X.x.-(y)zz, where "X.x".-are the initials of the name of the species; "y"-only for eukaryotes, indicates the cellular localization of the enzyme (c-cytosolic; m-mitochondrial; p-peroxisomal); "zz"-denotes the type of thiolase (kt-EC 2.3.1.16; at-EC 2.3.1.9; scpx-sterol carrier protein X/EC 2.3.1.16). The sequences used were:

- | | |
|---|---|
| 1) ORF P19-1 (this work); | 20) <i>R. norvegicus</i> mkt1 (Arakawa <i>et al.</i> , 1987); |
| 2) <i>A. thaliana</i> PKT3 (Hayashi <i>et al.</i> , 1998; Newman <i>et al.</i> , 1994); | 21) <i>H. sapiens</i> mtk2 (Kamijo <i>et al.</i> , 1994); |
| 3) <i>B. napus</i> pkt (Olesen <i>et al.</i> , 1997); | 22) <i>R. norvegicus</i> mkt2 (Kamijo <i>et al.</i> , 1993); |
| 4) <i>C. maxima</i> pkt (Kato <i>et al.</i> , 1996); | 23) <i>H. sapiens</i> scp-x (Ohba <i>et al.</i> , 1994); |
| 5) <i>C. sativus</i> pkt (Preisig-Müller & Kindl, 1993); | 24) <i>Mus musculus</i> scp-x (Seedorf <i>et al.</i> , 1993); |
| 6) <i>M. indica</i> pkt (Bojorquez & Gomez-Lim, 1995); | 25) <i>H. sapiens</i> mat (Fukao <i>et al.</i> , 1990); |
| 7) <i>H. sapiens</i> pkt (Bout <i>et al.</i> , 1988); | 26) <i>R. norvegicus</i> mat (Fukao <i>et al.</i> , 1989); |
| 8) <i>R. norvegicus</i> pktA (Bodnar & Rachubinski, 1990; Hijikata <i>et al.</i> , 1990); | 27) <i>C. tropicalis</i> pat (Kurihara <i>et al.</i> , 1992); |
| 9) <i>C. tropicalis</i> pktA (Kanayama <i>et al.</i> , 1994); | 28) <i>H. sapiens</i> cat (Song <i>et al.</i> , 1994); |
| 10) <i>S. cerevisiae</i> pkt (Igual <i>et al.</i> , 1991); | 29) <i>E. coli</i> at (Blattner <i>et al.</i> , 1997); |
| 11) <i>Y. lipolytica</i> pkt (Berninger <i>et al.</i> , 1993); | 29) <i>R. sativus</i> cat (Vollack & Bach, 1996); |
| 12) <i>Acinetobacter calcoaceticus</i> kt (Kowalchuk <i>et al.</i> , 1994); | 30) <i>S. cerevisiae</i> cat (Hiser <i>et al.</i> , 1994); |
| 13) <i>Alcaligenes eutrophus</i> kt (Peoples & Sinskey, 1989); | 31) <i>Schizosaccharomyces pombe</i> cat (Yoshioka <i>et al.</i> , 1997); |
| 14) <i>Alcaligenes sp.</i> kt (Lee <i>et al.</i> , 1997); | 32) <i>Caenorhabditis elegans</i> at (Sulston <i>et al.</i> , 1992); |
| 15) <i>Archaeoglobus fulgidus</i> kt (Klenk <i>et al.</i> , 1997); | 34) <i>Clostridium acetobutylicum</i> at (Petersen & Bennett, 1991); |
| 16) <i>E. coli</i> kt (DiRusso, 1990); | 35) <i>Haemophilus influenza</i> at (Tatusov <i>et al.</i> , 1996); |
| 17) <i>Pseudomonas fragii</i> kt (Sato <i>et al.</i> , 1992); | 36) <i>Rhizobium meliloti</i> at (Tombolini <i>et al.</i> , 1995); |
| 18) <i>P. putida</i> kt (Harwood <i>et al.</i> , 1994); | 37) <i>Thiocystis violacea</i> at (Liebergesell & Steinbuchel, 1993); |
| 19) <i>H. sapiens</i> mkt1 (Abe <i>et al.</i> , 1993); | 38) <i>Z. ramigera</i> at (Peoples <i>et al.</i> , 1987). |

The C-terminal 119 residues of the longer *H. sapiens* and *M. musculus* scpx sequences, lacking similarity to thiolases, are not shown. The *M. indica* sequence is known to be incomplete at the C-terminal (Bojorquez & Gomez-Lim, 1995).

		*	20	*	40	*	60	*	80	*	100	*	
1) PKT2	:	-----	MERAMERQKILLRHLNPVSSSSSLKHEPSLLSPVNCVSEVSP	----	MAAFGDDIIVIAAYETAICKARRGGFKITLPDDELLASVIRAVV	----	ERTSLDP	-----	:	93			
2) A.t.-PKT3	:	-----	MEKAIERQRVLLLEHLRP--SSSSSHNYEASLSASACLAGDSAAYQRTSLYGGDDVWIAAHRTPICKSKRCNFKOTYPDELLAPVIRALI	----	EKTNLNP	-----	:	94					
3) B.n.-pkt	:	-----	MEKAMERQRVLLLEHLRP--SSSSSHSFEGSLASACLAGDSAAYQRTSLYGGDDVWIAAHRTPICKSKRCNFKOTYPDELLAPVIRALI	----	EKTNLNP	-----	:	94					
4) C.m.-pkt	:	-----	MEKAINRQSILLHHLRP--SSSAYSHESSLSASVCAAGDSASYQRTSVFGDDVWIAAYETAICKSKRCGGFKOTYPDELLAPVIRALI	----	EKTNLNP	-----	:	93					
5) C.s.-pkt	:	-----	MEKAINRQSILLHHLRP--SSSAYTNESSLSASVCAAGDSASYQRTSVFGDDVWIAAYETAICKSKRCGGFKOTYPDELLAPVIRALI	----	EKTNLNP	-----	:	91					
6) M.i.-pkt	:	-----	MEKAINRQQVLLQHLRP--SNSSSHNYESALAASVCAAGDSAAYHRASVYGGDDVWIAAHRTPICKSKRCGGFKOTYPDELLAPVIRALI	----	EKTNLDP	-----	:	94					
7) H.s.-pkt	:	-----	MQRLQVVLGHLRGPADSGWMPQAAPCLSGAPQAS	----	AADVWVHGRRTPICRAGRGGFKITTPDELLSAVITAVL	----	KDVNLRP	-----	:	80			
8) R.n.-pktA	:	MSESVGRTSAMHRLQVVLGHLRGPRESSALQAAPCSATFPQAS	----	ASDVWVHGRRTPICRAGRGGFKITTPDELLSAVITAVL	----	QDVKLKP	-----	:	90				
9) C.t.-pktA	:	-----	MDRLNQLSGQLKPN	----	KQSILQKN	-----	PDDVWVIAAYETAICKGFKGCSFRSVRSEFILTEELKEFI	----	KKTNIDP	-----	:	69	
10) S.c.-pkt	:	-----	MSQRLQSIKDLVESAMGKGESKRKNSILLEKR	----	PEDVWVIAANPSAIGKGFKAFAKDVNTDYLLNFNEFIGFPEPLRADL	----	:	82					
11) Y.l.-pkt	:	-----	MDRLNNLATQLEQNP	----	AKGLDAITSKN	-----	PDDVWVIAAYETAICKGFKGCLFKDTSSSELLASLEGLV	----	KESKIDP	-----	:	72	
12) A.c.-kt	:	-----		----	MKHAYINDAIEPTPFGRYAGGLAAVRADLGAIPHAALIERNPSVN	----	:	45					
13) A.e.-kt	:	-----		----	MTDVWVISAARTAVG-KFGGSLAKIPAPELGAVVKAALERAGVKPE	----	:	46					
14) A.s.-kt	:	-----		----	MEVWVISAARTAVG-KFGGSLAKTPATELGAVVTEAIAEAGLSSD	----	:	46					
15) A.f.-kt	:	-----		----	MAVWVISAARTAVG-RFGGSLKVAAYELGAVVAKGMKKLGVPKPTPESIEFY	----	:	53					
16) E.c.-kt	:	-----		----	MEQVWVINDAIEPTMGRSKGGAFRNVRAEELSAHLASLLARNPALEA	----	:	47					
17) P.f.-kt	:	-----		----	MSLNPRDVWVDFGTPMGRSKGGMHRNTRAEDMSAHLTSKVLEENSKVDP	----	:	51					
18) P.p.-kt	:	-----		----	MRDVWVINDAIEPTPIGRFGGALAGVRADLAAVPAKALIEPNPAVQ	----	:	45					
19) H.s.-mkt1	:	-----		----	MRLLRGVFVIAAKETPFGRYAGGLLKDTATDLSEFAAKAALSAGKVSPE	----	:	49					
20) R.n.-mkt1	:	-----		----	MALLRGVFIWAAKETPFGRYAGGLLKDTATDLTEFAAKAALSAGKVPPE	----	:	49					
21) H.s.-mkt2	:	-----	MTILTYPFKNLPTASKWALRFSIRPLSCS-SQLRAAPAVQTKTKTLAKPNIRNVVVDGVVFPFLLS-GTSYKDLMPHBLARAATGLLHETSVPKEV	----	:	97							
22) R.n.-mkt2	:	-----	MTTILTSTFRNLSTTSKWALRFSVRPLSCS-SQVQSAPAVQTKSKTLAKPNLNKNIWVEGVIPFLLS-GTSYKDLMPHBLARAATSGLLYFTNIPKDV	----	:	98							
23) H.s.-scpx	:	-----		----	MSSSPWEPATLRRNFVWVGVMKFKVPGAENSRDYP--DLAEEAGKALADAQIP	----	:	53					
24) M.m.-scpx	:	-----		----	MPSVALKSPRLRRNFVWVGVMKFKMPGGENSRDYP--DMAKEAGQKALADAQIP	----	:	53					
25) H.s.-mat	:	-----	MAVLAALLRSGARSRSPLLRRLVQEIIRYVERSIVSKPTLKEVWVISAARTETPIG-SFLGSLSLPATKLGSLATQGAIEKAG--IP	----	:	82							
26) R.n.-mat	:	-----	MAALAVLHGTVRRPPLRGLLQEVRCIGRSYASKPTLNDVWVISAARTETPIG-SFLGSLASQPATKLGSLATQGAIEKAG--IP	----	:	79							
27) C.t.-pat	:	-----		----	MTLPPVYIVSTARTPIG-SFQGSLSLTYSDLGAAVRAALAKVPQIKP	----	:	48					
28) H.s.-cat	:	-----		----	MNAGSDPVVWVISAARTIIG-SFNICALAAVPVQDLGSTVKEVLEKATVAPE	----	:	50					
29) R.s.-cat	:	-----		----	MAHSADSSDNPRDVCIWGVARTPMG-GFLGSLSSLPATKLGSLATTAALKR--EMLT	----	:	54					
30) S.c.-cat	:	-----		----	MSQNVYIVSTARTPIG-SFQGSLSSTAVELGAVVAKGALAVPELDA	----	:	47					
31) S.p.-cat	:	-----		----	MVNTVEYVISAARTPMG-SFQGSFASLPATKLGSLATKGAIEKAG--NIK	----	:	47					
32) C.e.-at	:	-----	MLRAVSTSFGTARAASAV	----	AKKNMNPVWDAVTPPFVVS-GTVFKILMAVLLQKEAKALVEKTKLPYEQ	----	:	69					
33) E.c.-at	:	-----		----	MKNCVVISAARTIIG-SFNGLSLASTSAIDLGATVKAALIEKAKIDSQ	----	:	46					
34) C.a.-at	:	-----		----	MKEVVVISAARTIIG-SYGKSLKIVPAVLLGATATKEAVKAGIKPE	----	:	46					
35) H.i.-at	:	-----		----	MENVVWVISAARTPIG-SFNICALSSVSAVLLGAIVVQEAIKRANIESA	----	:	46					
36) R.m.-at	:	-----		----	MSNPSVWVISAARTAVG-SFNCAFGNTLAHELGAATKAVLEAGVEAG	----	:	48					
37) T.v.-at	:	-----		----	MSDTIVWVDAARTIIG-TFGGALSALQATLIGTTVNLALIEKGTGIAPE	----	:	47					
38) Z.r.-at	:	-----		----	MSTPSVWVISAARTAVG-SFNCAFANTPAHELGAIVVSAVLEAGVAAG	----	:	47					

```

120      *      140      *      160      *      180      *      200      *      220
1) PKT2      : -----SEVGDIVVGTVLAAGSQRASECMAAFYAFETVPLRVNRQSSQLQAVADVAARAFYDGGGAGVEEMTDH----- : 171
2) A.t.-PKT3 : -----SEVGDIVVGTVLAAGSQRASECMAAFYAFETVPLRVNRQSSQLQAVADVAARAFYDGGGAGVEEMTNP----- : 172
3) B.n.-pkt   : -----SDVGDIVVGTVLAAGSQRASECMAAFYAFETVPLRVNRQSSQLQAVADVAARAFYDGGGAGVEEMTNP----- : 172
4) C.m.-pkt   : -----SEVGDIVVGSVLAAGSQRASECMAAFYAFETVPLRVNRQSSQLQAVADVAARAFYDGGGAGVEEMVNE----- : 171
5) C.s.-pkt   : -----SEVGDIVVGSVLAAGSQRASECMAAFYAFETVPLRVNRQSSQLQAVADVAARAFYDGGGAGVEEMTNP----- : 169
6) M.i.-pkt   : -----SEVGDIVVGTVLAAGSQRASECMAAFYAFETVPLRVNRQSSQLQAVADVAARAFYDGGGAGVEEMTNP----- : 172
7) H.s.-pkt   : -----EQLGDI VCGNVLEPGAG-AIMAFIAQFLSDIETVPLRVNRQSSQLQAVADVAARAFYDGGGAGVEEMTNP----- : 154
8) R.n.-pktA  : -----EQLGDI VCGNVLEPGAG-AVMPFIAQFLSDIETVPLRVNRQSSQLQAVADVAARAFYDGGGAGVEEMTNP----- : 164
9) C.t.-pktA  : -----SLIEDVAIGNVLEQAAG-ATEHFGACLAAGIYTAAFIARFCSSQLMAISDIANKIKETIEIECGLAGGAEISMTNY----- : 146
10) S.c.-pkt  : -----NLIEEACGNVLENGAG-ATEHFGACLAAGIYSTPFVAFNRQSSQLMAINDIANKIKVQIDIGLALVESMTNNY----- : 159
11) Y.l.-pkt  : -----KLIGDVCGNVLAAGAG-ATEHFGACLAAGIETVPLRVNRQSSQLMAINDVANKIRAFQIDIGGCGVEEMTNP----- : 149
12) A.c.-kt   : -----WAQVDDVIYGCANQAGEDNRNVGMSAALLAGLEVVPATIRLQCGSSQLAMAAARAKAEAHIIAGEVEEMERAPYVMGKSEGA----F : 134
13) A.e.-kt   : -----QVSEVIMGVITAGSG-QNPAPQAAIKAGLEAMVPAMIRKVGCSQLKAMLAANAIMADAEVWAGGVEEMTNPAAEHVL-PGSRD----G : 131
14) A.s.-kt   : -----QIGEVIMGVITAGVG-QNPAPQASMKAGAKETPALIRKVGCSQLKAMLAQAAMWIDTEVWAGGVEEMTNPAAEHVL-PNSRD----G : 131
15) A.f.-kt   : EKLEKGRIDLESAYDYDGEVELEDEWIMGNVLAQAAG-QNPAPQASTLAKIKETIPAYIRKVGCSQLKALAYHAKAKDAKALIAGGVEEMTNPAPYAL-PKARW----G : 160
16) E.c.-kt   : -----AALDDVIYGCNQTTLEQGFNIAAALLAGIHSVPATIRLQCGSSQLAHDAARMIMTCDQAQCVGSEHGHVPMSHGVDFHP----G : 135
17) P.f.-kt   : -----GEMEDVIYGCNQTTLEQGNVIAASIMTQIHTSAAQVSRQCGSSQLHTAAQAMTONGDIFVVGVEEMTNPVMMHGVDPNP----H : 139
18) P.p.-kt   : -----WDQVDEVIYGCANQAGEDNRNVGMSAALLAGLEVVPATIRLQCGSSQLAMAAARAKAEAHIIAGEVEEMERAPYVMGKSEGA----F : 134
19) H.s.-mkt1 : -----TVDSVIMGNVLEQSSDAIYLAHVGRLVIRKETPALIRKVGCSQLFQSVNGCQELCVKFAEVLCCGTEEMTNPAPYCV-RNVRF----G : 135
20) R.n.-mkt1 : -----TVDSVIMGNVLEQSSDAAYLAHVGRLVIRKETPALIRKVGCSQLFQSVNGCQELCVKFAEVLCCGTEEMTNPAPYCV-RNVRF----G : 135
21) H.s.-mkt2 : -----VDYLIIFGTVLEQVKT-SNVAPQALGAFSDKTEPAHVTMACISNQATTAVGLIASQCDVWAGGVEEMTNPVIRHSRNMRLMLDLN : 187
22) R.n.-mkt2 : -----VDYLIIFGTVLEQVKT-SNVAPQALGAFSDKTEPAHVTMACISNQATTAVGLIASQCDVWAGGVEEMTNPVIRHSRNMRLMLDLN : 188
23) H.s.-scpx : -----YSAMDQACVGVYFGDSTCGQRAIYHSLGMTIIRIINNNNCATSTALFMAHQLOGLANCVLALGFERNERSIGT----- : 131
24) M.m.-scpx : -----YSAMDQACVGVYFGDSTCGQRAIYHSLGMTIIRIINNNNCATSTALFMAHQLOGLANCVLALGFERNERSIGT----- : 131
25) H.s.-mat   : -----KEEMKEAYMGNVLEQGGEG-QAPTEQAVLGAHISTPCTILNKVCASMKALMMASQSLMCHQDMVAGGVEEMTNPVYVM---NRG----S : 167
26) R.n.-mat   : -----KEEMKEAYMGNVLEQGGEG-QAPTEQAVLGAHISTPCTILNKVCASMKALMMASQSLMCHQDMVAGGVEEMTNPVYVM---NRG----S : 164
27) C.t.-pat   : -----QDVDEIVFGGVLEQNVG-QAPTEQVALKAGLEDSTIASTINKVCASMKALMMASQSLMCHQDMVAGGVEEMTNPVYVM---NRG----S : 134
28) H.s.-cat   : -----DVSEVIFGHVLAAGCG-QNPVROASVGAIEYSVPANSCMIGCSGLKAVCLAVQSIGIDSSVWAGGVEEMTNPVYVM---NRG----S : 134
29) R.s.-cat   : -----RLWSKEVFGNVLSANLG-QAPTEQVALGAGISNSVICTILNKVCASMKALMMASQSLMCHQDMVAGGVEEMTNPVYVM---NRG----S : 141
30) S.c.-cat   : -----SKDFDEIVFGNVLSANLG-QAPTEQVALGAGISNSVICTILNKVCASMKALMMASQSLMCHQDMVAGGVEEMTNPVYVM---NRG----S : 134
31) S.p.-cat   : -----SDVDEIVFGNVLSANLG-QNPAPQALGAFSDKTEPAHVTMACISNQATTAVGLIASQCDVWAGGVEEMTNPVIRHSRNMRLMLDLN : 187
32) C.e.-at    : -----LDHIIICGTVLEQCKT-SNIADEEALLAVEDKIPAHITLACISSNVANTTGMGMATGNANAIAGGVEEMTNPVIRHSRNMRLMLDLN : 159
33) E.c.-at    : -----HVDEVIMGNVLEQAGLG-QNPAPQALLKSLAETVCGFTVNVKCSGLKAVCLAVQSIGIDSSVWAGGVEEMTNPVYVM---NRG----S : 132
34) C.a.-at    : -----DVNEVILGNVLEQAGLG-QNPAPQALGAFSDKTEPAHVTMACISNQATTAVGLIASQCDVWAGGVEEMTNPVIRHSRNMRLMLDLN : 187
35) H.i.-at    : -----LVNEVIMGNVLEQAGLG-QNPAPQALGAFSDKTEPAHVTMACISNQATTAVGLIASQCDVWAGGVEEMTNPVIRHSRNMRLMLDLN : 187
36) R.m.-at    : -----EVDEVILGNVLEQAGLG-QNPAPQALGAFSDKTEPAHVTMACISNQATTAVGLIASQCDVWAGGVEEMTNPVIRHSRNMRLMLDLN : 187
37) T.v.-at    : -----QVSEVILGNVLEQAGLG-QNPAPQALGAFSDKTEPAHVTMACISNQATTAVGLIASQCDVWAGGVEEMTNPVIRHSRNMRLMLDLN : 187
38) Z.r.-at    : -----EVNEVILGNVLEQAGLG-QNPAPQALGAFSDKTEPAHVTMACISNQATTAVGLIASQCDVWAGGVEEMTNPVIRHSRNMRLMLDLN : 187

```

====Signature 1====

+++BLK A+++

		240	260	280	300	320	34
1) PKT2	:	--IPGGGFHGS--PRAQFPKARDCLLP-----	MGITSENVAEREGVTPLEEDMAAEESHKFAAATASEKFKDEIIPVATKIVDPETKAEKAIVSVLDGVFNPNADL	:	274		
2) A.t.-PKT3	:	--M---AWEGSVPAKKFAQAQNCLLP-----	MGVTSENVAORFGVSPLEEDQAAVDSHRFAAATAACKFKDEIIPVATKIVDPETKAEKAIVSVLDGIPPTTLIASL	:	273		
3) B.n.-pkt	:	--M---AWEGSVPAKKFEQAQNCLLP-----	MGVTSENVAORFGVSPLEEDQAAVDSHRFAAATAACKFKDEIIPVATKIVDPETKAEKAIVSVLDGIPPTTLIASL	:	273		
4) C.m.-pkt	:	--M---AWDGSVPRKSMENAQNCLLP-----	MGVTSENVAORFGVSPLEEDQAAVDSHRFAAATAACKFKDEIIPVATKIVDPETKAEKAIVSVLDGIPPTTLIASL	:	272		
5) C.s.-pkt	:	--M---AWEGSVPRKSIENAQNCLLP-----	MGVTSENVAORFGVSPLEEDQAAVDSHRFAAATAACKFKDEIIPVATKIVDPETKAEKAIVSVLDGIPPTTLIASL	:	269		
6) M.i.-pkt	:	--M---AWEGSVPRKKEFEQAQNCLLP-----	MGVTSENVAORFGVSPLEEDQAAVDSHRFAAATAACKFKDEIIPVATKIVDPETKAEKAIVSVLDGIPPTTLIASL	:	273		
7) H.s.-pkt	:	--LADRGNPGNITSRMEKEKARDCLIP-----	MGITSENVAEREGISPEKQITLALASQQAARASQKGCFAATIVPVTTTVHDDK-GTKRSITVTQDEGIRPSTTMEGL	:	257		
8) R.n.-pktA	:	--LSNRGNPGNISRLESKDARDCLIP-----	MGITSENVAEREGISPEKQITLALASQQAARASQKGCFAATIVPVTTTVHDDK-GTKRSITVTQDEGIRPSTTMEGL	:	267		
9) C.t.-pktA	:	--RD-PRVAPRIIPHLADAQMEKCLIP-----	MGITSENVAEREGISPEKQITLALASQQAARASQKGCFAATIVPVTTTVHDDK-GTKRSITVTQDEGIRPSTTMEGL	:	248		
10) S.c.-pkt	:	--KNVNPLGMSSEKQKNREAKKCLIP-----	MGITSENVAEREGISPEKQITLALASQQAARASQKGCFAATIVPVTTTVHDDK-GTKRSITVTQDEGIRPSTTMEGL	:	254		
11) Y.l.-pkt	:	--GP-NSVTP-FSNKFQNEEAKKCLIP-----	MGITSENVAEREGISPEKQITLALASQQAARASQKGCFAATIVPVTTTVHDDK-GTKRSITVTQDEGIRPSTTMEGL	:	250		
12) A.c.-kt	:	GRTQKIEDTTGWRFINPKKAMYGVDT-----	MPQTAENVAEQEGISPEKQITLALASQQAARASQKGCFAATIVPVTTTVHDDK-GTKRSITVTQDEGIRPSTTMEGL	:	235		
13) A.e.-kt	:	FRM---GDAKLVTMTVDGLWDVYNQYH-----	MGITAEENVAKQYGLTPEADEPAGVGNKAEAKKACKDEIIPVVL-----PQRKEDPVAFKTEFVRQATLDS	:	229		
14) A.s.-kt	:	QRM---GDWKIVTMTVDGLWDVYNQYH-----	MGITAEENVAKQYGLTPEADEPAGVGNKAEAKKACKDEIIPVVL-----PQRKEDPVAFKTEFVRQATLDS	:	229		
15) A.f.-kt	:	YRMSVTAKDEILMLVYDGLWEKFGYH-----	MGMTAENVIVERVIGISPEKQITLALASQQAARASQKGCFAATIVPVTTTVHDDK-GTKRSITVTQDEGIRPSTTMEGL	:	260		
16) E.c.-kt	:	LSRNVAKAAGH-----	MGITAEENVAKQYGLTPEADEPAGVGNKAEAKKACKDEIIPVVL-----PQRKEDPVAFKTEFVRQATLDS	:	219		
17) P.f.-kt	:	MSLYAAKASGH-----	MGITAEENVAKQYGLTPEADEPAGVGNKAEAKKACKDEIIPVVL-----PQRKEDPVAFKTEFVRQATLDS	:	223		
18) P.p.-kt	:	SRNMKLEDTTGWRFINPLMKSQYGVDS-----	MPETAENVADDYQVSPADQDAARASQQAARASQKGCFAATIVPVTTTVHDDK-GTKRSITVTQDEGIRPSTTMEGL	:	234		
19) H.s.-mkt1	:	TKLG---SDIKLESLWV-SLTDQHVQLP-----	MAMTAENVITVKHKISPEDECKYAAQSQQPKAANDAGYENDMANIEV-----KTKKQKQ-TQVDEHARPOLEQ	:	232		
20) R.n.-mkt1	:	TKFG---LDLKEETLWA-GLTDQHVQLP-----	MAMTAENVITVKHKISPEDECKYAAQSQQPKAANDAGYENDMANIEV-----KTKKQKQ-TQVDEHARPOLEQ	:	232		
21) H.s.-mkt2	:	KAKSMGQRLSLISKFRFNLAPELPAVSEFSTSETMGHSADRLAAAFVSELEODEAARHSLKKQDECHLS-DVVEFK	-----PQK---DTVTKDNGIRPSSLEQ	:	289		
22) R.n.-mkt2	:	KAKTLAQRLSLITKFRNLFLSPELPAVAEFSTNETMGHSADRLAAAFVSELEODEAARHSLKKQDECHLS-DVVEFK	-----PQK---DTVTKDNGIRPSSLEQ	:	290		
23) H.s.-scpx	:	----KFSDRTPTDKHVDLLINKYGLSAHPVAPQMFYAGKEHMEKYGTKIHFAGIKWKNHKHSVN-NPYSQODEY	-----LDEV	:	209		
24) M.m.-scpx	:	----KFSDRTPTDKHIEVLIDKYGLSAHPITPQMFYAGKEHMEKYGTKIHFAGIKWKNHKHSVN-NPYSQODEY	-----LDEV	:	209		
25) H.s.-mat	:	TPY---GGVKELIIVKGLTDVYNKIH-----	MGSCAENVAKKLNIAPNEODAAINITYTSKAEWEACKEGNEVIEVTV-----TV-KQPDVIVKEDEEYKRVDFSK	:	264		
26) R.n.-mat	:	TPY---GGVKELIIVKGLTDVYNKIH-----	MGSCAENVAKKLNIAPNEODAAINITYTSKAEWEACKEGNEVIEVTV-----TV-KQPDVIVKEDEEYKRVDFSK	:	261		
27) C.t.-pat	:	ARY---GDAVIVGVQKDGILDVYEEKL-----	MGVAAPKCAKDHGEFPEDEINEAINITYTSKAEWEACKEGNEVIEVTV-----TV-KQPDVIVKEDEEYKRVDFSK	:	232		
28) H.s.-cat	:	VKI---GEMPTTSLICDGLTDAFHNCH-----	MGITAEENVATKWQVSPEDQDKVAULSQNPTENAKAGHEDKEIVPVVL-----STRKGLIEKTDEFPFHGNEEAS	:	231		
29) R.s.-cat	:	SRF---GHDSLVGMLKGLWDVYNDG-----	MGSCAENVAKKLNIAPNEODAAINITYTSKAEWEACKEGNEVIEVTV-----TV-KQPDVIVKEDEEYKRVDFSK	:	239		
30) S.c.-cat	:	AKF---GQTVVGVVERDGLNDAYDGLA-----	MGVHAEKCARDNDITPEQDNDAAIESYQKSQKSQKEGKEDNEIVPVTV-----KGFRKPDQVTKDEEPARLHVKL	:	232		
31) S.p.-cat	:	AKY---GNVELVGLRDGLSDAYDGLP-----	MGNAEELCAEEHSDPASODAAISQYKFAQNRQATKAFEGEIVPVV-----PVGRKPNKVTDEEPPKNLNEDKL	:	231		
32) C.e.-at	:	KAKDVPSKLKGGQIVKNLSPELPAVAEFSTGETMGHSGDRLAAAFVSELEODEAARHSLKKQDECHLS-DVVEFK	-----DQKPKTIKEDNGIRVSTLEKL	:	263		
33) E.c.-at	:	YRL---GDGQYVILRDGLMCATHGYH-----	MGITAEENVAKQYGLTPEADEPAGVGNKAEAKKACKDEIIPVVL-----PQRKEDPVAFKTEFVRQATLDS	:	229		
34) C.a.-at	:	YRM---GNAKFVDEMIDGLWDAFNQCH-----	MGITAEENVAKQYGLTPEADEPAGVGNKAEAKKACKDEIIPVVL-----PQRKEDPVAFKTEFVRQATLDS	:	228		
35) H.i.-at	:	VKM---GNLTURDTMEDGLTCASNHYH-----	MGITAEENVAKQYGLTPEADEPAGVGNKAEAKKACKDEIIPVVL-----PQRKEDPVAFKTEFVRQATLDS	:	229		
36) R.m.-at	:	VKM---GDYKIDTMDKGLTDAFYGYH-----	MGITAEENVAKQYGLTPEADEPAGVGNKAEAKKACKDEIIPVVL-----PQRKEDPVAFKTEFVRQATLDS	:	229		
37) T.v.-at	:	QRM---GDWPMKIDTMDKGLTDAFYGYH-----	MGITAEENVAKQYGLTPEADEPAGVGNKAEAKKACKDEIIPVVL-----PQRKEDPVAFKTEFVRQATLDS	:	230		
38) Z.r.-at	:	VKM---GDFKIDTMDKGLTDAFYGYH-----	MGITAEENVAKQYGLTPEADEPAGVGNKAEAKKACKDEIIPVVL-----PQRKEDPVAFKTEFVRQATLDS	:	227		

	0	360	380	400	420	440	
1) PKT2	: AKLPTV	-KQN-ESTTAGNAGQSDGAGAVLMKRSIMKKKPIIG	-----VFRSPA	VTGVEI-SWGLG	-----EAVNTPAATKIA	-GPNVSD	ILFEETEFASQYVYS : 372
2) A.t.-PKT3	: GKLPV	-KKD-GTTTAGNAGQSDGAGAVLMKRSVAMQKPIVG	-----VFRTPAAVGVDE	-AIGI	-----EAVNTPAAVKAE	-GLEDDI	ILFEETEFASQYVYC : 371
3) B.n.-pkt	: GKLPV	-KKD-GTTTAGNAGQSDGAGAVLMRRSVATQKPIVG	-----VFRTEAAVGVDE	-AIGV	-----EAVNTPAAVKAE	-GLEDDI	ILFEETEFASQYVYC : 371
4) C.m.-pkt	: GKLPV	-KKD-GTTTAGNAGQSDGAGAVLMKRSVAMQKPIIG	-----VFRTEAVGVDE	-AIGV	-----EAVNTPAAVKAE	-GLEDDI	ILFEETEFASQYVYC : 370
5) C.s.-pkt	: GKLPV	-KKD-GTTTAGNAGQSDGAGAVLMKRSVAMRKPIIG	-----VFRTEAVGVDE	-AIGV	-----EAVNTPAAVKAE	-GLEDDI	ILFEETEFASQYVYC : 367
6) M.i.-pkt	: AKLPV	-KKD-GTTTAGNAGQSDGAGAVLMKRSVAMRKPIIG	-----VFRTEAAVGVDE	-AIGV	-----EAVNTPAAVKAE	-GLEDDI	ILFEETEFASQYVYC : 371
7) H.s.-pkt	: AKLPV	-KKD-GTTTAGNAGQSDGAGAVLMRRSKAEELPIIG	-----VFRSYAVVGVP	-DILG	-----EAYNTPVALQK	-GLTVS	WULFEETEFASQAAYC : 355
8) R.n.-pktA	: AKLPV	-KDG-ESTTAGNAGQSDGAGAVLMRRSKAEELPIIG	-----VFRSYAVVGVP	-DILG	-----EAYNTPAAVQK	-GLTVN	WULFEETEFASQAAYC : 365
9) C.t.-pktA	: GKLPV	---D-GTTTAGNAGQSDGAGAVLMKRSIAEAKYPIIG	-----KYVLCSTAGVP	-ELIG	-----EAYNTPVVKRT	-GLTVS	WULFEETEFASQAAYC : 344
10) S.c.-pkt	: SSIPV	-IKDRGTTTAGNAGQSDGAGAVLMRRSVANQLNPIVG	-----RYIDFQTVGVPE	-ELIG	-----EAYNTPKVLKAT	-GLTVS	WULFEETEFASQAAYC : 353
11) Y.l.-pkt	: GKLPV	-SAE-GTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----KRVHCKTVGVPE	-ELIG	-----EAYNTPAVLEDL	-GLTVN	WULFEETEFASQAAYC : 348
12) A.c.-kt	: AKLPV	-VKPEAVLGGAGGNDGAAVLASDEAVAQYQKARA	-----KRIASTTVGIEP	-RILGFA	-----EAPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 333
13) A.e.-kt	: SGIPV	-DKAGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----TIKSYANAGVDE	-KVLGMS	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 327
14) A.s.-kt	: AGIPV	-DKAGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----RTAAGTSGIDE	-ATIG	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 327
15) A.f.-kt	: AKLPV	-KKD-GTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----RILAVASAGID	-AYIG	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 358
16) E.c.-kt	: ATIPV	-DPVNMVITAGTSSALSDGAAVLMSSESAHELKPIIA	-----RVRSMVAVGCD	-STIG	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 318
17) P.f.-kt	: AAIPV	-NPKGVAVAGTSSALSDGAAVLMSSESAHELKPIIA	-----VIRSMVAVGCD	-STIG	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 322
18) P.p.-kt	: TKLPV	-NGPDKIVLGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----RVLGMSAGVDE	-RVIG	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 332
19) H.s.-mkt1	: QKLPV	-KKD-GTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----RIVGYFVSGCD	-STIG	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 330
20) R.n.-mkt1	: QKLPV	-KKEGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----RVVGYFVSGCD	-AIGIG	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 330
21) H.s.-mkt2	: ANIPV	-I-KPYGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----YLRDEMYVSQDEKDOILLG	-----TYNTPKVLK	-----GLTLD	WULFEETEFASQAAYC : 389	
22) R.n.-mkt2	: ANIPV	-I-KPYGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----YLRDEMYVSQDEKDOILLG	-----TYNTPKVLK	-----GLTLD	WULFEETEFASQAAYC : 390	
23) H.s.-scpx	: MASIPV	-----DFLILQCCPTSDGAGAVLMRRSVAEKLPPIIA	-----VEILAQEMMTDLPSSFEESIKVMGFDMSKE	-ARKCYEKS	-----GLTLD	WULFEETEFASQAAYC : 315	
24) M.m.-scpx	: MRSIPV	-----DFLILQCCPTSDGAGAVLMRRSVAEKLPPIIA	-----VEILAQEMMTDLPSTFEESIKVMGFDMSKE	-ARRCYEKS	-----GLTLD	WULFEETEFASQAAYC : 315	
25) H.s.-mat	: PKLPV	-QKENGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----RIVAFADAVER	-IDFPIA	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 363
26) R.n.-mat	: PKLPV	-QKENGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----RIVAFADAVER	-IDFPIA	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 360
27) C.t.-pat	: KSLIPV	-QKENGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----KISGWGEAART	-FDFPIA	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 331
28) H.s.-cat	: SKLPV	-LTDGTGIVIPAGGNDGAAVLMSSESAHELKPIIA	-----RIVSWSQVGVPE	-SINGIG	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 331
29) R.s.-cat	: RKLIPV	-KENGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----KVKGYGDAQVE	-EFFTTA	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 340
30) S.c.-cat	: RSIPV	-QKENGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----IKGWGEAAHQ	-ADFTWA	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 332
31) S.p.-cat	: KSLIPV	-KSNGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----KIKGWGEAAHQ	-ERFTTS	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 329
32) C.e.-at	: SSIPV	-KPHGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----YLRDLYLVAQD	-KDOILLG	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 363
33) E.c.-at	: GKLIPV	-DKAGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----RIKSYASGVPE	-AIGIG	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 327
34) C.a.-at	: AKLPV	-KKD-GTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----KIVSYGSAGVD	-AIGIG	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 326
35) H.i.-at	: AKLPV	-KKEGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----KIRSYASGVDE	-SVGLG	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 327
36) R.m.-at	: AKLPV	-DKEGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----RIVSWATAGVD	-QIGTIC	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 327
37) T.v.-at	: GKLIPV	-DKAGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----RIVAFSSAGVD	-AIGIG	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 328
38) Z.r.-at	: AKLPV	-DKEGTTTAGNAGQSDGAGAVLMRRSVAEKLPPIIA	-----RIVSWATAGVD	-KVLGTC	-----EVPNPKKLKQ	-GLTLD	WULFEETEFASQAAYC : 325

++BLK B++

++++BLK C++

	460	*	480	*	500	*	520	*	540	*	560	
1) PKT2	: CKKE	-----	LDMEK	-----	NVNGGAI	IGHPLGATGACVATLLHEMKRRGKDCRFVISMICITETMAAAVFERGDSVDNLSNARVANGDSH	-----					: 457
2) A.t.-PKT3	: RNKLG	-----	LDPEK	-----	NVNGGAMA	IGHPLGATGACVATLLHEMKRRGKDCRFVISMICITETMAAAVFERGDGVDELNRNARKVE-AQGLLSKDAR	-----					: 462
3) B.n.-pkt	: RNKLG	-----	LDPEK	-----	NVNGGAMA	IGHPLGATGACVATLLHEMKRRGKDCRFVISMICITETMAAAVFERGDGVDELNRNARKVE-SHGLLSKDAR	-----					: 462
4) C.m.-pkt	: RNKLG	-----	LDPEK	-----	NVNGGAI	IGHPLGATGACVATLLHEMKRRGKDCRFVISMICITETMAAAVFERGDCVDELCAKKVE-GINLLSKDAR	-----					: 461
5) C.s.-pkt	: RNKLG	-----	LDPEK	-----	NVNGGAI	IGHPLGATGACVATLLHEMKRRGKDCRFVISMICITETMAAAVFERGDCVDELCAKKVEGGVNLLSKDAR	-----					: 459
6) M.i.-pkt	: RNKLG	-----	LDPEK	-----	NVNGGAMA	IGHPLGATGACVATLLHEMKRRGDCRFVISMICITAPRA	-----					: 430
7) H.s.-pkt	: VEKLR	-----	LDPEK	-----	NPLGGAVAL	IGHPLGCTGACQVITLLNELKRRGKRA-YVWSMISITETMAAAVFEYPGN	-----					: 424
8) R.n.-pktA	: VEKLR	-----	LDPEK	-----	NPLGGAVAL	IGHPLGCTGACQVITLLNELKRRGKRA-YVWSMISITETMAAAVFEYPGN	-----					: 434
9) C.t.-pktA	: AEQVN	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 408
10) S.c.-pkt	: IHKLG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 417
11) Y.l.-pkt	: IQHCG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 414
12) A.c.-kt	: TRDLG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 401
13) A.e.-kt	: HQQMG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 392
14) A.s.-kt	: NKELG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 392
15) A.f.-kt	: IKELG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 424
16) E.c.-kt	: IKDLG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 387
17) P.f.-kt	: LKDLK	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 391
18) P.p.-kt	: LRELK	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 400
19) H.s.-mkt1	: ERSID	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 397
20) R.n.-mkt1	: QKSLD	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 397
21) H.s.-mkt2	: FKANDSDWFAENYMGRKTKVGLPPLKLF	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 474
22) R.n.-mkt2	: FKANDSDWFAENYMGRKTKVGLPPLKLF	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 475
23) H.s.-scpx	: YEALGLCPEGQGATLVDRGNTYGGKWMINPSGELISKGHPLGATGLAQCAELCWQLPGEAGKRQVPGAKALQHNLEIGGAVVVTLYKMGFPEAASSFRTHQIEAVPTSSAS											: 428
24) M.m.-scpx	: YEALGLCPEGQGATLVDRGNTYGGKWMINPSGELISKGHPLGATGLAQCAELCWQLPGEAGKRQVPGAKALQHNLEIGGAVVVTLYRMGFPEAASSFRTHQVSAAPTSSAG											: 428
25) H.s.-mat	: IKMLE	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 427
26) R.n.-mat	: IKMLE	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 424
27) C.t.-pat	: AELMK	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 403
28) H.s.-cat	: VKELG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 397
29) R.s.-cat	: QKLLG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 406
30) S.c.-cat	: TKILK	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 398
31) S.p.-cat	: TKILG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 395
32) C.e.-at	: LNA	DSDFCKEQMKRSGKFGMPMDKL	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----				: 448
33) E.c.-at	: GKNLG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 394
34) C.a.-at	: AKDLK	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 392
35) H.i.-at	: GKDLN	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 393
36) R.m.-at	: NKDLG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 393
37) T.v.-at	: NQELG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 394
38) Z.r.-at	: NKDLG	-----	LDPEK	-----	NVNGGAI	IGHPLGCTGACQYATLLPLKPG--Q--IGTSMISITETMAAAVFEYPGN	-----					: 391

====Signature 2====

==Signature 3=

Signature 1

(LIVM)(NST)XXC(SAGLI)(ST)(SAG)(LIVMFYNS)X(STAG)(LIVM)XXXXXX(LIVM)

90/133 VNRQCSSGLQAVADVAASI**Signature 2**

NXXGGX(LIVM)(SA)XGHPXGX(ST)G

341/384 NVNGGAIAIGHPLGATG**Signature 3**

(AG)(LIVMA)(STAGLIVM)(STAG)(LIVMA)CX(AG)X(AG)X(AG)X(SAG)

378/421 GVISMCIGTGMGAA**Figure 7.6** Thiolase PROSITE signature sequences found in P18-1 and P19-1.

The signature consensus are shown together with their occurrences in the two ORFs. Alternative aminoacids at a given position are enclosed in brackets. The positions of the initial residues of each in P18-1/P19-1 are indicated. In bold typeface and underlined are the two cysteine and the histidine residues known to be catalytically active in thiolases (Masamune *et al.*, 1989; Palmer *et al.*, 1991; Thompson *et al.*, 1989; Williams *et al.*, 1992).

the sequence of P19-1 is aligned (GAP method) with that of the *S. cerevisiae* peroxisomal thiolase (Igual *et al.*, 1992) and where the elements of secondary structure experimentally identified in the latter (Mathieu *et al.*, 1994; 1997) or predicted for P19-1 are indicated. P18-1 was subjected to identical analysis by the same methods to determine if the absence of the 43 additional residues of P19-1 would have an effect on the secondary structure at its N-terminus. The predictions were virtually identical to those obtained with P19-1 with only negligible differences being registered for some residues near the N-terminus (data not shown). These did not alter the consensus prediction relative to that obtained for P19-1.

Some of the elements found in the yeast thiolase are predicted by all tested methods to occur in P19-1, although with different lengths. Examples include the expected α -helices for the segment DDLLASVLKAVVERTS, and the β -sheet for the segment DIVVGTV, at positions 76 and 98, respectively. Because modelling by the different methods relies on distinct types of information there is an increased likelihood for the existence of these elements in P19-1. Other elements of secondary structure found in the yeast protein are expected to be present in P19-1 according to some methods, but not others, and vice-versa. For instances, for the well conserved segment IGIGAGVESMST at position 158, only

SOPMA predicts the same alternation of elements as is found in the yeast thiolase, consisting of a β -sheet followed by an α -helix. All other methods fail to predict an α -helix in the region. Limited β -sheets are projected for this segment by most other methods while the DPM method only ascribes such conformation to a single residue. On the other hand, α -helices were expected in the segments AYRTAICKA at position 56 and GAVLLMK at position 300. The equivalent segments in the yeast enzyme assume a different structure. Overall, many of the elements of secondary structure found in the yeast thiolase enzyme, were predicted to be present in P19-1 by a majority of the methods, as is indicated by the calculated consensus structure (Fig. 7.7). Yet, significantly, this general observation does not extend to the two cysteine residues Cys₁₂₇ and Cys₄₂₆ and immediate surrounding segments, in spite of their very high degree of conservation.

Current methods for secondary structure prediction have limited accuracy. The estimated accuracies for three-state predictions of the set of methods employed in this study range from 62.2% for the homology method (Levin *et al.*, 1986), 64.3% for the GOR method (Garnier *et al.*, 1996), 65.0% for PREDATOR (Frishman & Argos, 1996), 69.5% with SOPMA (Geourjon & Deleage, 1994; 1995), up to 71.4% for the PHD method (Rost *et al.*, 1994). A common limitation to most methods, except PREDATOR, is their inherent inability to directly incorporate the contribution to its secondary structure of long range interactions between residues brought together by the folding of a protein. Inevitably questions arise on how this drawback, and indeed other methodological faults, affected the correspondence between the observed secondary structure of the *S. cerevisiae* thiolase and that predicted for P19-1. To obtain some information in this regard, the *S. cerevisiae* peroxisomal 3-ketoacyl-CoA thiolase sequence was subjected to the same secondary structure prediction methods. For proteins sharing a significant degree of similarity (see Table 7.3), it was expected that the methods would generate analogous models if they had homologous structures. Remarkably, the various models obtained for the yeast thiolase found substantially better correspondence to those of P19-1 than did the experimentally determined secondary structure of the protein. Worthwhile examples are found in the segments containing the two catalytically important cysteine residues for which the derived consensus secondary structures are very close to those obtained for P19-1. A graphic display of the consensus derived from the modelling of the yeast thiolase sequence and that of P19-1 is shown in Figure 7.8. These results, taking in consideration the known limitations in accuracy of the various prediction methods, are compatible with the notion that P19-1 has a similar structure to that of the *S. cerevisiae* peroxisomal enzyme.

Figure 7.7 Prediction of secondary structure elements of P19-1 and comparison with the elements found in the *S. cerevisiae* peroxisomal thiolase.

The sequence of P19-1 was subjected to analysis by a variety of methods (Section 2.11). The methods included: **PHD**) a neural-network method (Rost & Sander, 1993, 1994; Rost *et al.*, 1994), performed at EMBL (Heidelberg); **GOR4**) an improved GOR statistical method (Garnier *et al.*, 1996; Gibrat *et al.*, 1987); **HM**) the homology method of Levin *et al.* (1986); **DPM**) the Double Prediction Method (Deleage & Roux, 1987a, 1987b); **PREDATOR**) a method based on estimation of non-local interactions (Frishman & Argos, 1996), **SOPMA**) the Self-Optimized Method from Alignments (Geourjon & Deleage, 1994; 1995), all performed at I.B.M.P. (France). The sequence of P19-1 is shown aligned with that of the *S. cerevisiae* peroxisomal thiolase (*S.c.*), as performed by GAP (cf. Fig. 7.3). For each method, residues involved in forming helices (**H**) and β -sheets (**E**) are indicated. Residues predicted to be part of coils (**c**), turns (**t**) or loops (**l**) are also shown. A consensus secondary sequence profile for three states was calculated for each position. A minimum of 50% of matches to a given state were required for it to be assigned as consensus. For consensus calculation c, t, and l were considered as a same coil (c) state. In the consensus, residues with an ambiguous status are indicated by a question mark (?). For PHD, only the residues with an average probability of accuracy of >82% are shown. No predictions are given for the other residues as indicated by the dashes. The helices and β -sheets found in the yeast enzyme (Mathieu *et al.*, 1994, 1997) are also shown on the top line, above its sequence. These correspond to the elements found in chain B of the dimer as described in the PDBsum database (Laskowski *et al.*, 1997). The cysteine residues involved in catalysis in the yeast enzyme, and the corresponding ones in P19-1 are shown in bold typeface (and underlined).


```

      .           .           .           .           .
S.c.    1 .....MSQRLQSIKDHLVESAMGKGESKRKNSLLEKRPEDVIVAANR 43
      :| . | :|.. .| :...: .. | :. .:|:| | | | |
P19-1   1 MERAMERQKILLRHLNPVSSSSSLKHE..PSLLSPVNCVSEVSPMAAFGDDIVIVAAYR 58

PHD     1-1-HHHHHHHHHHHH---1111-111..1111111111-111---111-EEEEEE--
GOR4    cccccHHHHHHHHHcccccccccccccc..ccEEccccccccccccccccEEEEHHHH
HM      HHHHHHHtEEcEEccccccccHcccHc..ccccctccccctHccHcHctctEEEEcccc
DPM     ccHHHHHHHHHHHcccccttttccccct..ccccccEccEcHEccHHHHccccEEEEHHHH
PREDATOR cccHHHHHHHHHHHcccccccccccccc..ccccccccccccccccccccEEEEHHHH
SOPMA   HHHHHHccccEEEcccccccccccHHHcc..HHHHHHccccEEccccEEccccEEEEEEEE
Consensus cHHHHHHHHHHHcccccccccccccc..ccccccccccccccccccccEEEEHHHH

```

```

      .           .           .           .           .
S.c.    44 SAIGKGFKGAFKDVNTDYLLYNFLNEFTGRFPEPLRADNLIEEVACGNVLNVGAG.ATE 102
      •||.: |:| | | . | || ..|...:| . . . | . :...:|.:|..| . | |
P19-1   59 TAICKARRGGFKDTLPDDLASVLKAVVERTS...LDPSEVGDIVVGTVIAPGSQRAME 114

PHD     --1---111---11--HHHHHHHHHHHHHH-1...11----EEEEEE--11111-H
GOR4    HHHHHccccccccccccHHHHHHHHHHHHHcc...cccccccEEEEEEccccHHHH
HM      HcccHcctccccctcccHHHHHHHHHHHHcccc...cccccccEEEEEEccccHHHH
DPM     HHHHHHccctccccccccHHHHEHHHEHHHcc...cccccccEEEEEEccccHHHHH
PREDATOR HHHHHHccccccccccccHHHHHHHHHHHHHc...cccccccEEEEEEccccHHHHH
SOPMA   EccEEccccccccccccHHHHHHHHHHHHHHcc...ccHHHcccEEEEEEccccHHHHH
Consensus HHHHHHccccccccccccHHHHHHHHHHHHHcc...cccccccEEEEEEcccc?HHHH

```

```

      .           .           .           .           .
S.c.    103 HRAACLASGIPYSTPFVALNRQCSGLTAVNDIANKIKVGQIDIGLALGVESMTNNYKNV 162
      |.|. . |:| |. .:| | | | | | | | | | | | | | | | | | | | | | | |
P19-1   115 CRVAAYFAGFPDSVPVRTVNRQCSGLQAVADVAASIRAGYYDIGIGAGVESMSTDHPIG 174

PHD     HHHHHHH-111111-----HHHHHHHHHHHHHHHHHH-11--EEE-11---111111
GOR4    HHHHEEEccccccccEEEEccccccHHHHHHHHHHHHHccEEEEccccEEEEcccccc
HM      HHHHHHEHccccccccccEccccctHcHHHHHHHHHHHccccccEEccccccccctcctcct
DPM     HHHHHHHHccccccccEEEEcccttcHHHHHHHHHHHHccccccccctccEccccccccct
PREDATOR HHHHHHHHccccccccEEEEccccccHHHHHHHHHHHHHccccEEEEcccccccccccc
SOPMA   HHHHHHHHccccccccEEEEcHHHHHHHHHHHHHHHHHHccccEEEEEEEEHHHEEccc
Consensus HHHHHHH?cccccccEEEEcccccc?HHHHHHHHHHHHcccc?EE?ccc?cccccccccc

```

```

      .           .           .           .           .
S.c.    163 NPLGMISSEELQKNREAKKCLIPMGITNENVAANFKISRKDQDEFAANSYQKAYKAKNEG 222
      .: . |.. |. .:|.:| | | | | | | | | | | | | | | | | | | | | |
P19-1   175 GGFH.GSNPRAQDFPKARDCLLPMGITSENVAERFGVTREEQDMAAVESHKRAAAAIASG 233

PHD     111-.-1---11-----1---HHHHHH-111-HHHHHHHHHHHHHHHHHHH-1
GOR4    cccc.cccccccccccccccEEccccccccHHHHHcCHHHHHHHHHHHHHHHHHHHHH
HM      cccc.cccccccccccHccccccccccctHcHHHHHccccctHHHHHHHHHHHHHHHHHt
DPM     ccct.tctccccHcCHHcHHccccccHHHHHHHcEHHHHHHHHHHHHHHHHHHHHcc
PREDATOR cccc.cccccccccccccEEccccccHHHHHHHccccHHHHHHHHHHHHHHHHHHHc
SOPMA   cccc.cccccccccccccEEccEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHc
Consensus cccc.cccccccccccccccE?ccccccHHHHHHHcc?HHHHHHHHHHHHHHHHHHHc

```

```

      .       .       .       .       .
      E       E       HHHHH
S.c.  223  LFEDEILPI.....KLPDGSICQSDEGPREPNVTAESLSSIRPAFIKDRGTTTAGNA 273
      :|||:|:|       | :. :. |:.||| . ..|.:.:..| |:.|.||||
P19-1 234  KLKDEIIPVATKIVDPETKAEKAIVVSVDGVRPNSNMADLAKLKTVF.KQNGSTTAGNA 292

PHD      1-1--E--EEEE--111111--EEEE-111111-HHHHHHHHHHH-1111----11
GOR4     cccccEEcEEEEccccHHHHHHHEEEEEccccccccHHHHHHHHHHHEE.Ecccccccc
HM        cccHHHcHHHccccctccHcHHHEEEEEccccctcHHHHHHHHHHHH.ctttcccccc
DPM       cHcHEEEcEEcEEcHcHHHHHEEEEEccccctcHHHHHHHHHHHEH.tttccccctc
PREDATOR  cccccEEEcEEEEccccHHHHHHHEEEEEccccccccHHHHHHHHHHHc.ccccccccc
SOPMA     ccccHHccEEEEEccccHHHHHHHEEEccccccccHHHHHHHHHccc.cccccEEEc
Consensus cccc?EE?EEEEccccccHHHHHEEEEEccccccccHHHHHHHHHHHc.ccccccccc

```

```

      .       .       .       .       .
      EEEEEEEEEHHHHHH      EEEEEEEEE HHHHHHHHHHHHHHHHHHH
S.c.  274  SQVSDGVAGVLLARRSVANQLNLPVLGRYIDFQTVGVPPEIMGVGPAYAI PKVLEATGLQ 333
      ||:||||.:||| :||:| . .||:| : .|...|||.|.:||:| ||| . . .||.
P19-1 293  SQISDGAGAVLLMKRSLAMKGLPILGVFRSFAVTGVEPSVMGIGPAVAIPAATKLAGLN 352

PHD      -1--HHHHHHHHHHHHHHHH1111--EEEE--E-1111-111111-HHHHHHHHH-11-
GOR4     cccccHHHHHHHHHHHHHHHccccccEEEEccccccccccccccccHHHHHHHHHccc
HM        cccctctcHEEEHHHHHHctttccEEEEEEEEEEcEccccEEcEccccEEcHHHHHHccc
DPM       cccctccHHHHHHHHHHHHccccccECEEHEEEHccEcccEcccccccHcHHcHHHccc
PREDATOR  cccccHHHHHHHHHHHHHHccccccEEEcEEccccccccccccEEHHHHHHHHccc
SOPMA     cHHHcccHEEEccccCHccccHHHHHHHHHEEEccccEEcccccccEEccccEEEc
Consensus cccccHHHHHHHHHHHHHccccccEEEEEE?cccccc?cccccc??H?HHHHHHccc

```

```

      .       .       .       .       .
      HHH  EEEE  HHHHHHHHHHH  HHH  HHHH  HHHHHHHHHHH
S.c.  334  VQDIDIFEINEAFAAQALYCIHKLGLIDLNKVNPRGGAIALGHPLGCTGARQVATILRELK 393
      | |||:||||||| | :|:..||:|:|:| |..| |||:| ||| | |||:|:| |
P19-1 353  VSDIDLFEINEAFASQYVYSCKKLELDMEKVNNGGATAIGHPLGATGARCVATLLHEMK 412

PHD      --11---HHHHHHHHHHHHHH--11111--1111-EE--111--HHHHHHHHHHHH
GOR4     cccHHHHHHHHHHHHHHHHHccEEEEccccccccEEEEccccccccHHHHHHHHHH
HM        cctccEcEccHtHHHHHEEEttccEEEtctccctcHEEEccccctcHHHHHHHHHH
DPM       cccccHHHHHHHHHHHcEEcccHHHHHHHHHEtttccccccccctcctHHHEHHHHHH
PREDATOR  ccccEEHHHHHHHHHHHcHHHHHHHHHHHEEEccccEEEEccccccccHHHHHHHHHH
SOPMA     cccccEEHHHHHHHHccccccccHHHHHHHHHEEHccccEEcHHHHcCHHEEEHHHH
Consensus cccccE?HHHHHHHHHH?cHc?HHHHH??ccccccccEE?ccccccHHHHHHHHHH

```

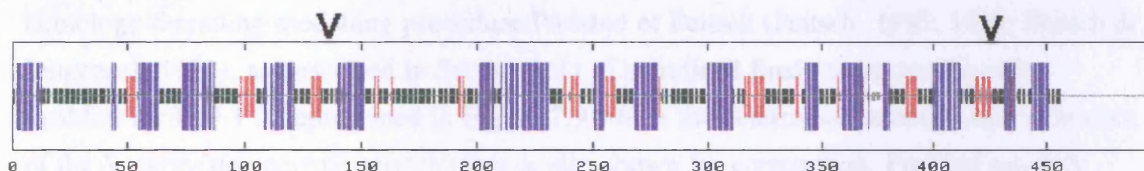
```

      .       .       .       .
      EEEEEEE EEEEEEE
S.c.  394  KD....QIGVSMIGTGMGAAAFIKE..... 417
      : .||:||||||| ||:| :
P19-1 413  RRGKDCRFVISMIGTGMGAAAVFERGDSVDNLSNARVANGDSH 457

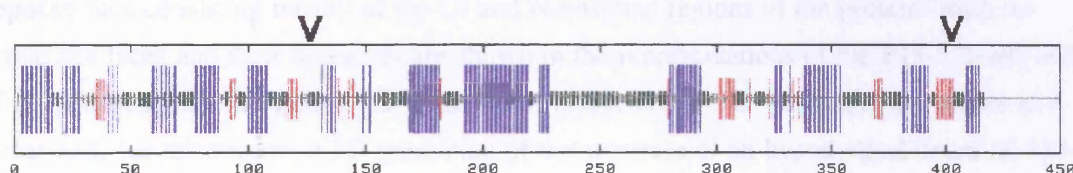
PHD      H-11111--EEEE--111--EEEE1111-----11111
GOR4     HccccEEEEEEEEEEccccHHHHHHccccccHHHHHEEEcEEEc
HM        HHccccEcEEcEEccccccHHHEHctccHHHHHHHHHHtcccc
DPM       HctcctcEcEEEEEEcctcHHHHHHHcctccccccHHHHtcccc
PREDATOR  HccccEEEEEEEEccccccccEEEEcccccccccccccccc
SOPMA     HcccccccEEccccEEEEHHHHHHHHccccHHHHHHHHHHHH
Consensus HcccccccEEEEEE?ccccHHH?HHcccccc??H?HHHHcccc

```

A) P19-1 model



B) Yeast model



C) Yeast (experimental)

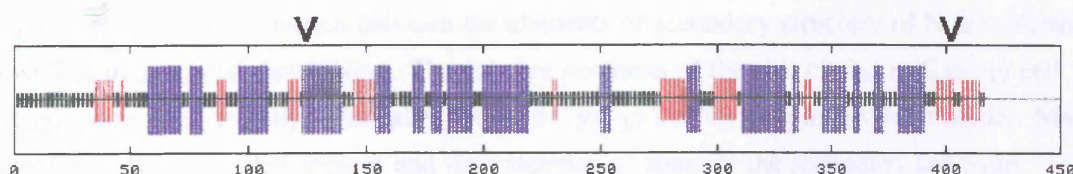


Figure 7.8 Secondary structure models for P19-1 and *S.cerevisiae* thiolase.

Shown are diagrammatic representations of the consensus secondary structures obtained for: **A)** P19-1; **B)** *S. cerevisiae* peroxisomal 3-ketoacyl-CoA thiolase. The consensus structures were obtained as described in Figure 7.7 except that the predictions obtained with the PHD method were not considered. **C)** interpretation of the secondary structure of the experimentally determined crystal structure of the *S. cerevisiae* enzyme (Mathieu *et al.*, 1994; 1997). The secondary structure interpretation was obtained from the PDBsum database (Laskowski *et al.*, 1997). Diagrams A) and B) were obtained from I.B.C.P. (France) and diagram C) was kindly prepared by Dr. Deleage (France).

7.7 Construction of a 3D model for P19-1/P18-1

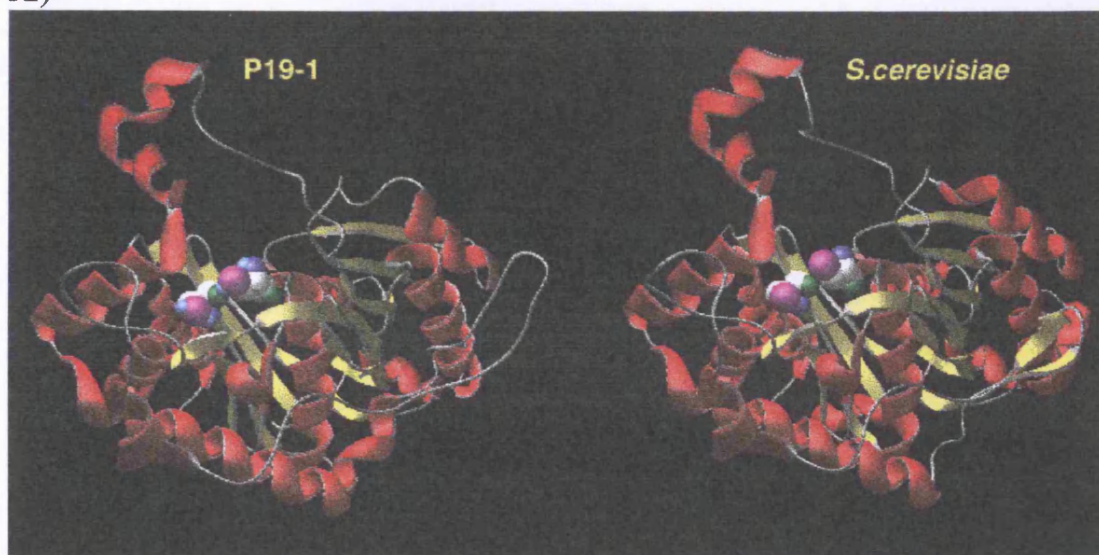
The availability of a crystallographic description for the *S. cerevisiae* thiolase (Mathieu *et al.*, 1994; 1997) and its degree of similarity to P19-1 (and P18-1) allowed its use as a template for homology-based modelling of the three-dimensional structure of these putative proteins to a high degree of confidence. To this end the P18-1 and P19-1 sequences were subjected to the homology-threading modelling procedure ProMod of Peitsch (Peitsch, 1995; 1996; Peitsch & Jongeneel, 1993), as described in Section 2.11. The refined final model developed by ProMod for P19-1 is represented in Figure 7.9 where the determined structure of a monomer of the *S. cerevisiae* peroxisomal thiolase is also shown for comparison. ProMod can only initiate modelling from the point where sequence similarity begins. For this reason, the N-terminal 43 residues of P19-1 were omitted from the models, as were the 2 C-terminal residues (cf. alignment in Fig. 7.4). This effectively made the two models identical as the N-extension of P19-1 relative to P18-1 is 43 residues-long. As described by Mathieu (Mathieu *et al.*, 1994), the active site pockets are located on one of the faces of the thiolase monomer, the opposite face consisting mostly of the C- and N-terminal regions of the protein. Both the active site faces and their opposites are shown in the representations of the P19-1 model and of the yeast enzyme in Figures 7.9A and 7.9B, respectively. The yeast enzyme occurs as a dimer and, for this reason, a 3D prediction of the structure of an hypothetical dimer of P19-1 was also constructed in a similar manner to that of the monomer (Section 2.11). This involved modelling of each of the monomers on a different subunit of the yeast thiolase and subsequent assembling of the two. In the dimer, both active sites are exposed on the same side (Mathieu *et al.*, 1994), as can be seen in Figure 7.9C for the modelled dimer of P19-1.

The most striking feature of the 3D-model of P19-1 monomer was its overall structural similarity to the *S. cerevisiae* enzyme as is evident in Figures 7.9A and 7.9B. There is nearly complete coincidence between the elements of secondary structure of both molecules as well as their spatial distribution. The relative positions of the side chains of Cys₁₃₇ and Cys₄₂₆ were also virtually identical to those of Cys₁₃₇ and Cys₄₂₆ in the yeast model. Most exceptions affected coiled regions and the extension of some of the secondary structure elements. The latter, identified in the stereoscopic Figures 7.10A and 7.10B, include extensions to common elements (α -helices H1, H3 and H4), and a few only present in the yeast enzyme (α -helix H2, and β -sheets S1 and S2). All these differences are located in the periphery of the molecules, away from the catalytically important cysteines (Fig. 7.10). The smallest distance to these cysteines was the 15 Å separating the α -helix extension H3 in the yeast thiolase and its Cys₁₂₅ (Cys₁₃₇ in P19-1). The other apparently closer element in Figure 7.10A, α -helix H2, is located 30 Å from the cysteines, behind the plane of the paper. Most differences between the two models affecting the coiled regions occur exclusively at or near the surface of the molecules (Fig. 7.10C).

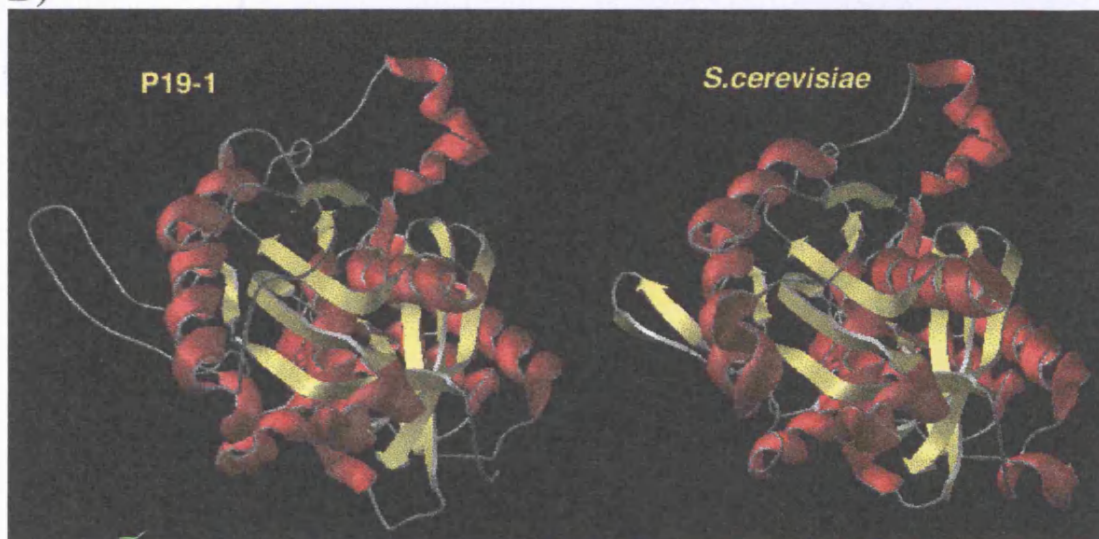
Figure 7.9 Ribbon models of the homology-threaded prediction of the 3D structure of P19-1 (P18-1).

- A)** Views of the P19-1 monomer and of one of the *S. cerevisiae* templates (corresponding to the chain B of the peroxisomal EC 2.3.1.16; Mathieu *et al.*, 1997). The ribbon representations were constructed with Swiss-PdbViewer (Guex & Peitsch, 1997). Helices are shown in red; Sheets are colored yellow; Coiled regions are shown as a grey thread. The side chains of the catalytically important Cys₁₂₅ and Cys₄₀₃ of the yeast enzyme, and the corresponding Cys₁₃₇ and Cys₄₂₆ of P19-1 are exposed on the portrayed face of the monomers where they are visible as space-filled atoms. All atoms are shown for the two P19-1 cysteines. For clearer view, hydrogen atoms are omitted from the yeast thiolase Cys₁₂₅ and Cys₄₀₃. Atom colors are: Carbon-white; Hydrogen-cyan; Nitrogen-blue; Sulphur-violet.
- B)** View of the opposite face of the molecules in A).
- C)** View of a ribbon model of an hypothetical dimer of P19-1. The face exposing the side chains of Cys₁₃₇ and Cys₄₂₆ is shown. One of the chains (chain A) is uniformly colored to distinguish it from the other (chain B). The 3D predictions of the P19-1 monomer and dimer were obtained as described in Section 2.11.

A)



B)



C)

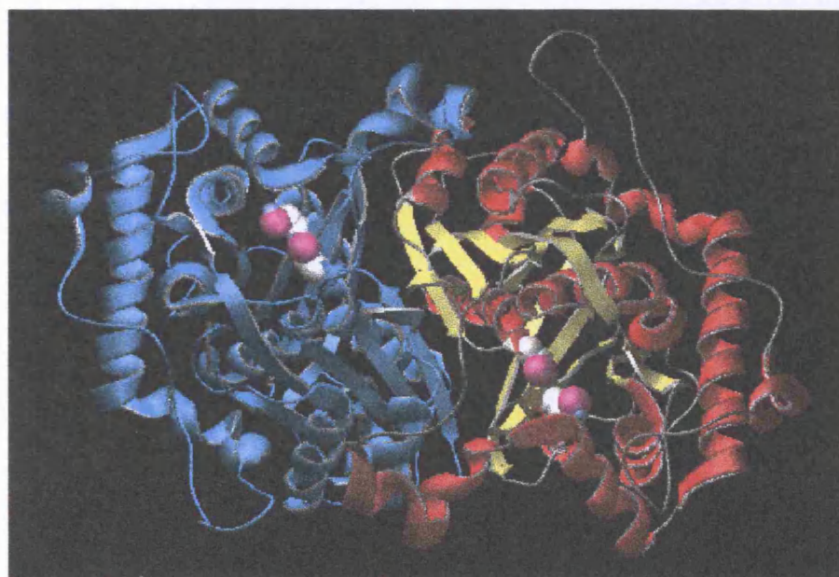


Figure 7.10 Visualization of the structural differences between the 3D prediction for P19-1 and the *S. cerevisiae* peroxisomal 3-ketoacyl-CoA thiolase.

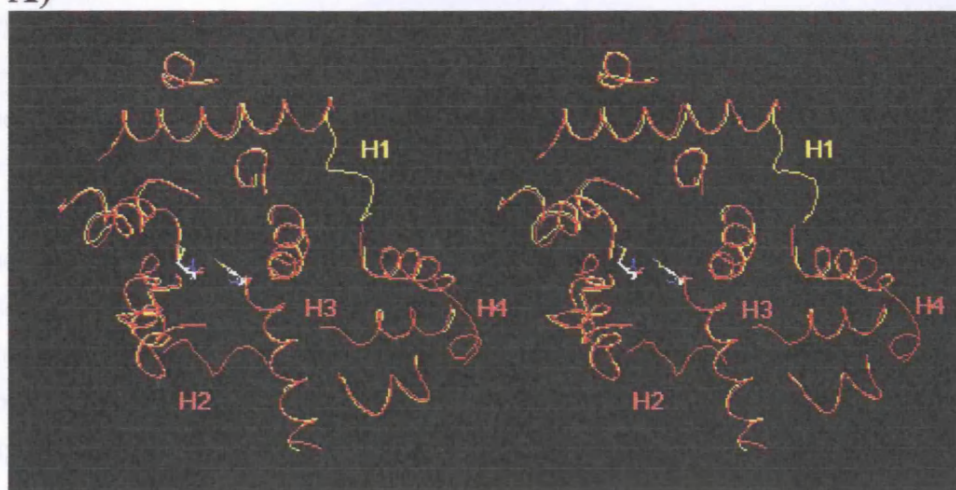
The stereoscopic figures show superpositions of the structural elements of P19-1 (**Yellow**) and the yeast thiolase (**Red**). Represented as single thread ribbons are the peptide backbones of the structural elements. Also shown as wireframe are the side chains of the two catalytically significant cysteines (Cys₁₂₅ and Cys₄₀₃) of the yeast enzyme, and the corresponding cysteine residues of P19-1 (Cys₁₃₇ and Cys₄₂₆). The ribbons models were obtained with Swiss-PdbViewer (Guex & Peitsch, 1997).

A) Helices; Significant differences are labelled with H1-H4.

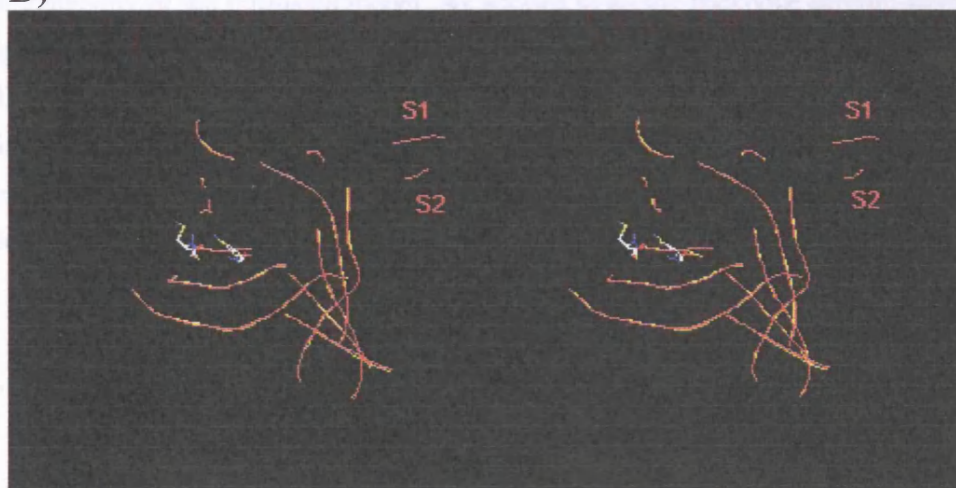
B) β -sheets; The two extra sheets in the yeast thiolase are indicated by S1 and S2.

C) Coiled regions.

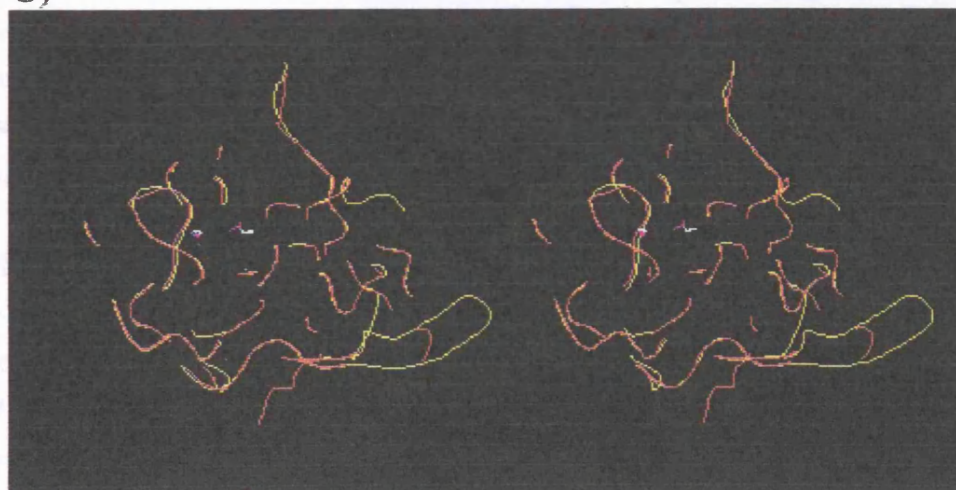
A)



B)



C)



The first concern in the construction of 3D-models is that of their reliability as representations of the actual structure that is subjacent to them. ProMod assigns a confidence value, the so-called B-factor, to the modelled positioning of each atom that essentially measures its deviation from the template structures. The closer the threading of a sequence brings it to the folding of its template, the higher the confidence associated with the model thus generated. The B-factors are arbitrarily defined as ranging from 0 to 100 the value increasing in direct proportion with the level of uncertainty. For the P19-1 model, the associated B-factors were generally quite low (12.5-25.0), *i.e.* it had a high degree of confidence. This is shown in Figure 7.11 where the B-factors are correlated with temperature colors. Only a few short segments, labelled A-G (colored red), were associated with the highest B-factor (99.9). These corresponded to segments of low similarity to the template structures as can be seen in the pairwise alignment in Figure 7.13 where all overlap gaps. The root mean square (RMS) of the sum of the distances separating equivalent atoms in the predicted and the reference structures constitutes another measure of their similarity and of the reliability of the estimated model. As applied to all the backbone carbon, oxygen and nitrogen atoms in the model the observed RMS value is of only 0.77 Å. Figure 7.12 depicts a 3D model temperature-colored for RMS values. It is clear from the picture that the segments displaying the highest RMS values (shown in red) coincide or are contained within the segments associated with maximum B-factor values. Where this occurs, the segments are identified as in Figure 7.11.

7.8 Cellular localization of the putative P18-1 and P19-1 thiolases

7.8.1 P19-1 is a putative peroxisomal protein (PKT2)

The high structural similarity of P18-1 and P19-1 to 3-ketoacyl-CoA thiolases leaves little doubt about their belonging to this group of enzymes (see above). As indicated earlier, 3-ketoacyl-CoA thiolases occur in mitochondria and peroxisomes of eukaryotes. Import of proteins, and particularly of thiolases, into either of these cellular compartments is dependent on appropriate sorting signals. Therefore, to gain insight into the possible localization of the two proteins P18-1 and P19-1, their sequences were analyzed to detect putative targeting signals.

Clearly, their greater similarity to peroxisomal rather than to mitochondrial thiolases (Section 7.4) suggests that P18-1 and P19-1 are peroxisomal enzymes. In rat (Swinkels *et al.*, 1991) and *S. cerevisiae* (Glover *et al.*, 1994) import of thiolases (EC 2.3.1.16) into peroxisomes, as well as of a minority of other enzymes (Subramani, 1993) is dependent on an N-terminal targeting signal. A consensus sequence for this peroxisomal targeting signal, designated PTS2, was originally derived by de Hoop and Ap (de Hoop & Geert, 1992), and,

Figure 7.11 Confidence B-factor indicator for the P19-1 3D prediction.

The 3D model of P19-1 is shown temperature-colored according to the B-factor values associated with its backbone atoms. B-factors ranged from 12.5 (Dark-blue) to 99.9 (Red). The highest (red) values correspond to short segments, labelled A to G, comprising all the reconstructed residues which could not be aligned with the template yeast thiolase structures. Only the backbone atoms are represented except for the Cys₁₃₇ and Cys₄₂₆ for which the side chains are also shown as space-filled models.

Figure 7.12 RMS deviations of P19-3 3D prediction (relative to templates).

The RMS deviations were calculated for the C, O and N backbone atoms. A temperature-colored model was constructed where brighter colors correspond to higher RMS values. The segments with highest deviations (red) are labelled as the corresponding segments with highest B-factors in Figure 7.11. Only the backbone atoms are represented except for the Cys₁₃₇ and Cys₄₂₆ for which the side chains are also shown as space-filled models.

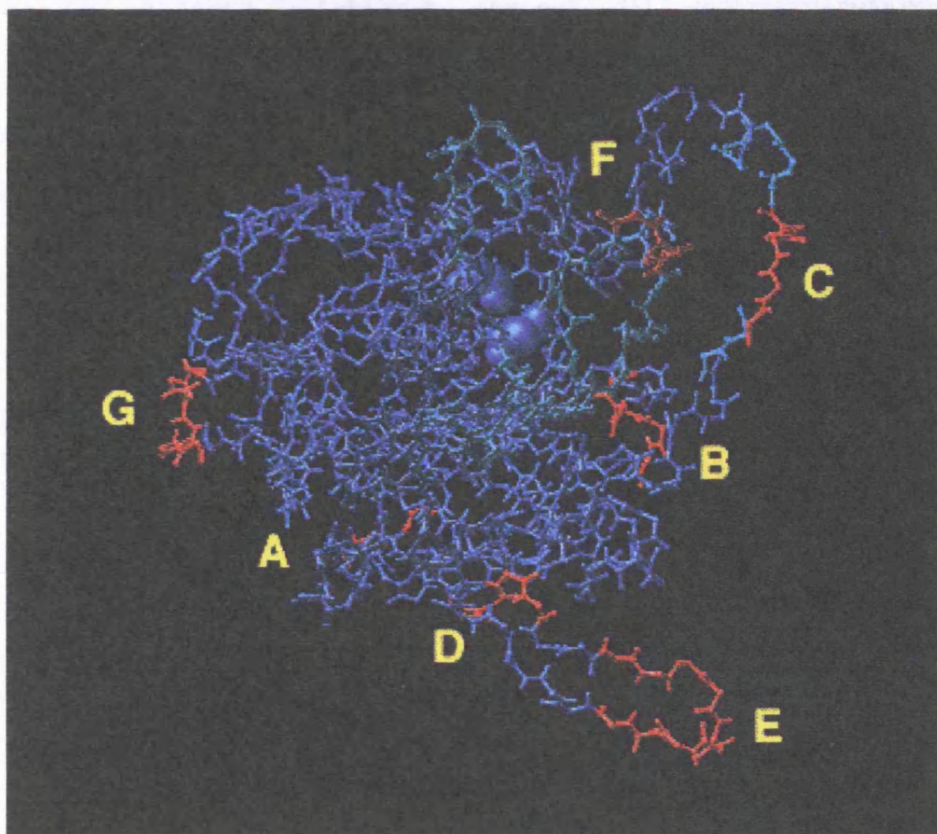


Figure 7.11

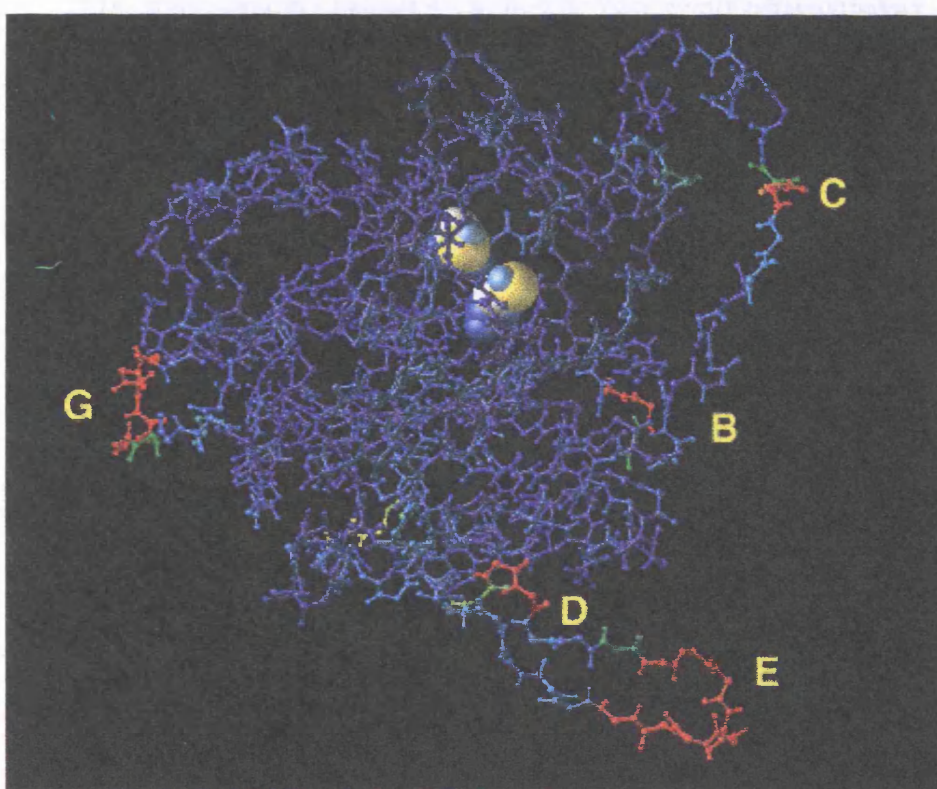


Figure 7.12

```

P19-1      1  MERAMERQKILLRHLNPNVSSNSSLKHE--PSLLSPVNCVSEVSPMAAFG 48
              :| . |:. . .| :...: . . | :. .
S.c.       1  MSQRLQSIKDHLVESAMGKGESKRKNSLLEKRP 33

P19-1      49 DDIVIVAAYRTAICKARRGGFKDTLPDDLIASVLKAVVERTS----LDPS 94
              :|:||||| |.|||. :|:||||. .| || .|. . .:| . . | .
S.c.       34 EDVVIVAANRSAIGKGFGKAFKDVNTDYLLYNFLNEFIGRFPEPLRADLN 83

P19-1      95 EVGDIVVGTVIAPGSQRAMECRVAAYFAGFPDSVPVRTVNRQCSSGLQAV 144
              :...:|. |:. . . . | | |. . . . :| |. |. . :|:|:|:|:| | |
S.c.       84 LIEEVACGNVLNVGAG-ATEHRAACLASGIPYSTPFVALNRQCSSGLTAV 132

P19-1     145 ADVAASIRAGYYDIGIGAGVESMSTDHIPGGGFH-GSNPRAQDFPKARDC 193
              .|:|. . :|. |. |||:| | | | . . . . . :| . . |. . :|:|
S.c.     133 NDIANKIKVGQIDIGLALGVESMTNNYKNVNPLGMISSEELQKNREAKKC 182

P19-1     194 LLPMGITSENVAERFGVTRREEQDMAAVESHKRAAAAIASGKLKDEIIPVA 243
              |:|:|:|:|:|:|:|:| . . :|. || |. |. . :| | . . | :|:|:|:|
S.c.     183 LIPMGITNENVAANFKISRKDQDEFAANSYQKAYKAKNEGLFEDEILPI- 231

P19-1     244 TKIVDPETKAERKAIVVSVDGVRPNSNMADLAKLKTVF-KONGSTTAGNA 292
              | . . :. . |:|. || |. . .|:|:|:|:| |:|. |. |||:|
S.c.     232 -----KLPDGSICQSDEGPRPNVTAESLSSIRPAFIKDRGTTTAGNA 273

P19-1     293 SQISDGAGAVLLMKRSLAMKKGLPILGVFRSFAVTGVEPSVMGIGPAVAI 342
              ||:|:|:|. :|:|:| . . |:|:| : . |. . .|:|. |:|:|:| | |
S.c.     274 SQVSDGVAGVLLARRSVANQLNLPVLGRYIDFQTVGVPPEIMGVGPAYAI 323

P19-1     343 PAATKLAGLNVSDIDLFEINEAFASQYVYSCKKLELDMEKVN NVNGGAIAI 392
              | . . .|:| | | |:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|
S.c.     324 PKVLEATGLQVQDIDIFEINEAFAAQALYCIHKLGLDLNKNV NPRGGAIAL 373

P19-1     393 GHPLGATGARC VATLLHEMKRRGKDCRFGVISMCI GTGMGAAAVFERGDS 442
              |||:|:|. | | |:|:|:|:| . :|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|
S.c.     374 GHPLGCTGARQVATILRELKGD---QIGVVSMCI GTGMGAAAFI KE 417

```

Figure 7.13 Segments of P19-1 with associated low confidence values in 3D modelling. The segments are identified by double overlining in a GAP alignment of P19-1 and the *S. cerevisiae* peroxisomal thiolase EC 2.3.1.16. The GAP alignment was obtained as described in Figure 7.4.

since, several variants have been put forward (Gietl, 1996; Subramani, 1996). However, none of these suggested consensus encompasses all known targeting sequences. A compilation of the various published consensi indicates that PTS2 is (RK)-(LVIQ)-X₅-(HQ)-(LA), as shown in Figure 7.14A. Matches to this consensus are found in all known or putative peroxisomal thiolases described to date including those of man (Bout *et al.*, 1988), rat (Bodnar & Rachubinski, 1990; Hijikata *et al.*, 1990), and the plants *C. maxima* (Kato *et al.*, 1996), *C. sativus* (Preisig-Müller & Kindl, 1993), *M. indica* (Bojorquez & Gomez-Lim, 1995), *B. napus* (Olesen *et al.*, 1997), *S. cerevisiae* and *Y. lipolytica* (Swinkels *et al.*, 1991), and *C. albicans* (Kanayama *et al.*, 1994), (Fig. 7.14B).

P19-1 exhibited the sequence RQKILLRHL six residues downstream from its putative N-terminus, fully conforming to the proposed PTS2 consensus and particularly similar to those found in the other plant enzymes at equivalent positions (Fig. 7.14B). Furthermore, in higher eukaryotes, PTS2 is within a presequence that is cleaved following import of the protein into the peroxisome, as has been demonstrated for watermelon (*Citrullus lanatus*; Gietl, 1990) and rat (*R. norvegicus*; Hijikata *et al.*, 1990) thiolases, and other enzymes including the glyoxysomal citrate synthase of *C. maxima* (Kato *et al.*, 1995) and the watermelon (Gietl, 1990) and *C. maxima* glyoxysomal malate dehydrogenases (unpublished, cited in Kato *et al.*, 1995). For these enzymes hydrolysis occurs near a cysteine residue located 24-43 residues from their N-termini, generating either Cys-COOH or Cys-Xxx-COOH (Kato *et al.*, 1995). A cysteine residue was present, at an equivalent distance (Cys₃₇) from the N-terminus in P19-1 (Fig. 7.14B). The putative *Arabidopsis* P19-1 protein thus displayed the recognized attributes of a peroxisomal 3-ketoacyl-CoA thiolase and was for this reason re-named PKT2. The numbering (2) being used to distinguish its gene from *PKT1*, the homologue encoding P18-1 (see below).

7.8.2 Absence of targeting signals in P18-1(PKT1)

Inspection of the P18-1 sequence showed that the absence of the 43 residues-long N-terminal extension relative to P19-1 results in its lack of a PTS2 signal. 3-ketoacyl-CoA thiolases are found only in two cellular compartments in eukaryotes, namely the peroxisome and the mitochondria. Import of most proteins into the peroxisomal matrix relies on a carboxy-terminal signal, PTS1. To date, no known cases have been reported of thiolases possessing such a signal. Several variations of PTS1 have been identified (reviewed in Subramani, 1993), which can be summarized into the consensus (SAC)-(KRHS)-(LMFI). The signal is active only when placed at the C-terminus. Neither P18-1 nor P19-1 display a match to this consensus at their carboxy ends.

Given the lack of any obvious peroxisomal import signal in P18-1 its sequence was inspected for the presence of other organellar targeting signals, namely mitochondrial. Thiolases of the EC 2.3.1.16-type occur in mitochondria of eukaryotes *e.g.* human (Abe *et*

A

(De Hoop & Geert, 1992): R-X₅-(HQ)-L
 (Subramani, 1993): (RK)-(IVL)-X₅-(HQ)-(LA)
 (Gietl, 1996): R-(QIL)-X₅-(HQ)-L
Consensus: (RK)-(LIVQ)-X₅-(HQ)-(LA)

B

* * *

```

P19-1      : -----MERAMERQKILLRHLNPVSSSNSSLKHEPSLLSPVNCVSEVSPM
A.t. PKT3 : -----MEKAIERQRVLLHLRPSSSSSHNYEASLSASACLAGDSAAYQR
B.n. PKT   : -----MEKAMERQRVLLHLRPSSSSSHSFEGSLASACLAGDSAAYQR
M.i. PKT   : -----MEKAINRQQVLLHLRPSSSSSHNYESALAASVCAAGDSAAYHR
C.m. PKT   : -----MEKAINRQSILLHLRPSSSAYSHESSLASVCAAGDSASYQRT
C.s. PKT   : -----MEKAINRQSILLHLRPSSSAYTNESSLSASVCAAGDSASYQRT
R.n. PKTA : MSESVGRTSAMHRLQVVLHLLAGRPESSSALQAAPCSATFPQASASDVVV
R.n. PKTB : -----MHRLQVVLHLLAGRSESSSALQAAPCSAGFPQASASDVVV
H.s. PKT   : -----MQRLQVVLHLRGPADSGWMPQAAPCLSGAPQASAADVIV
S.c. PKT   : -----MSQRLQSIKDHLVESAMGKGESKRKNSLLEKRPEDVVIVAA
C.t. PKT   : -----MDRLNQLSGQLKPNAQSILQKNPDDVVIVAAYRTAIGKG
C.t. PKT   : -----MDRLNQLSGQLKPNAQSILQKNPDDVVIVAAYRTAIGKG
Y.l. PKT   : -----MDRLNNLATQLEQNPAKGLDAITSKNPDDVVITAAYRTAH
  
```

== ==

Figure 7.14 Occurrence of PTS2 signals in P19-1.

- A) Consensus PTS2 signal from compiled published consensus.
 B) N-termini of P19-1 and peroxisomal EC 2.3.1.16 enzymes aligned at their matches to the PTS2 signal. Highlighted are the PTS2 matches and the cysteines known or presumed to be near the sites of cleavage of the pre-sequence in higher eukaryotes. Referencing for the various sequences is the same as in Figure 7.5.

al., 1993) and rat (Arakawa *et al.*, 1987). Although none have been identified in plant mitochondria there is controversy in relation to their existence (see Chapter 11). Most nuclear-encoded mitochondrial proteins depend on an N-terminal targeting signal of 20-60 residues including abundant positive charges and hydroxylated residues but few negative charges (Neupert, 1997). Six of the first 40 residues of P18-1 were arginines and lysines and no negative charges were found in it. Two negatively charged and one positively charged residues occurred in the following 20 positions. Only three hydroxylated residues were present in the segment up to position 40 and three more up to position 60. The P18-1 sequence was also submitted to analysis for detection of targeting signals with a knowledge-based interpretation tool, PSORT (Nakai & Kanehisa, 1992). No signals were detected by this program.

In spite of the lack of evidence with regard to its cellular site of activity, because of its homology to PKT2, the putative protein P18-1 was renamed peroxisomal 3-ketoacyl-CoA thiolase 1, or PKT1. The suffix 1 was chosen as the 5' end of the *PKT1* transcripts mapped upstream from those of the *PKT2* transcripts.

7.8.3 Determining the localization of PKT1 and PKT2

The unusual structure of PKT1 made it of interest to probe its intracellular localization. A project was initiated to achieve this purpose and simultaneously to confirm the presumed peroxisomal targeting of PKT2. The near structural identity between the two putative proteins posed the major obstacle to any such project, particularly if the N-terminus of PKT2 is proteolytically processed upon import into the peroxisome. To obviate this problem an epitope-tagging strategy was devised. It consisted of transforming *Arabidopsis* with a gene encoding a tag-modified PKT1 (or PKT2) under the control of a strong promoter. This would generate structurally distinct proteins and simultaneously allow their specific detection by established protocols. The epitope tag chosen (EQKLISEEDLN), derived from the human *c-myc* protein, was known to give a good immunohistochemical signal and not to interfere with the function of the tagged protein (Munro & Pelham, 1987). In addition, a well-studied monoclonal antibody against the *c-myc* epitope was available that does not cross-react with other cellular proteins (Evan *et al.*, 1985).

Three stages were envisaged to obtain suitable constructs for stable transformation of *Arabidopsis*. Firstly, a segment encoding *c-myc* was to replace the 3' end of the inserts in pCEN1-8 and pCEN1-9. In the second and third stages the modified pCEN-x inserts were to be placed downstream from a 35S promoter and the ensemble moved to a plant transformation vector. As reported here, only the first stage was completed.

To place the *c-myc* fragment at the end of the two proteins the segment between the downstream *Pst* I site in the inserts of the cDNA clones and the *EcoR* I site at the polylinker was replaced with a short *EcoR* I/*Pst* I DNA fragment encoding the epitope. The fragment

was obtained by annealing of two deoxyoligonucleotides, prMYC-F and prMYC-R, as described in Section 2.5.6. A number of manipulations were required to obtain the desired constructs due to the restriction maps of the cDNA clones (Section 2.5.1; Fig. 7.15):

- Step 1. Digestion of pCEN1-9 with *Sac* I and *Eco*R I. Separation of the fragments by electrophoresis. Isolation and purification of the downstream *Sac* I/*Eco*R I insert fragment containing the 3' terminal 0.32 kb of the cDNA.
- Step 2. Cloning of the 0.32 kb *Sac* I/*Eco*R I insert fragment into pBluescript. Plasmid pSKSE-3 was obtained.
- Step 3. Digestion of pSKSE-3 with *Pst* I and *Eco*R I, followed by agarose gel separation of the resulting small and large fragments. The small fragment was discarded and the linearized large fragment containing the vector sequences and the *Sac* I/*Pst* I cDNA segment was purified.
- Step 4. Ligation of the purified large fragment from step 3 with the *c-myc* fragment and cloning. Plasmid pSKMYC-1 was isolated.
- Step 5. Inversion of the orientations of the inserts in pCEN1-8 and pCEN1-9 by digestion with *Eco*R I and re-ligation followed by cloning in *E. coli*. One clone for each were selected and designated pCINV1-8 and pCINV1-9.
- Step 6. Digestion of pCINV1-8 and pCINV1-9 with *Sac* I, followed by gel separation and purification of the 1.2 kb insert-derived fragments.
- Step 7. Ligation of the cDNA *Sac* I fragments into previously *Sac* I-linearized pSKMYC-1, followed by transformation of *E. coli*. Resulting plasmids were designated pC18MYC and pC19MYC.

The structures of the plasmids obtained for the various constructs were confirmed by restriction enzyme analysis. Plasmid pSKSE-3 was digested with *Eco*R V plus *Sac* I which digest on each side of the 321 bp insert releasing it (Fig. 5.16A, lane 1). An insert-derived fragment shorter by 101 bp was expected to be generated by digestion with *Eco*R V plus *Pst* I, and was indeed observed (Fig 7.16A, lane 2). Digestion of pSKMYC-1 with *Pst* I should linearize the plasmid whereas combined digestion with *Pst* I plus *Eco*R V was expected to remove a small (46 bp) fragment including *c-myc*. The size reduction was (barely) visible upon separation of the products of the digestions in the gel shown in Figure 7.16B.

Figure 7.15 Construction of *c-myc* tagged derivatives of pCEN1-8 and pCEN1-9.

A schematic representation of various steps involved in replacing the 3' ends of the cDNAs with the *c-myc* tag is shown. Detailed descriptions of the various steps are given in the text. Vector sequences are shown as thin lines and cDNA segments as thick lines. The direction of the ORFs is indicated by the arrows. Relevant restriction enzyme sites are labelled as: *EcoR* I (E), *Pst* I (P) and *Sst* I (S). The internal *EcoR* I site at 47 bp from the 5' end of the cDNA in pCEN1-9 is not indicated. Not drawn to scale.

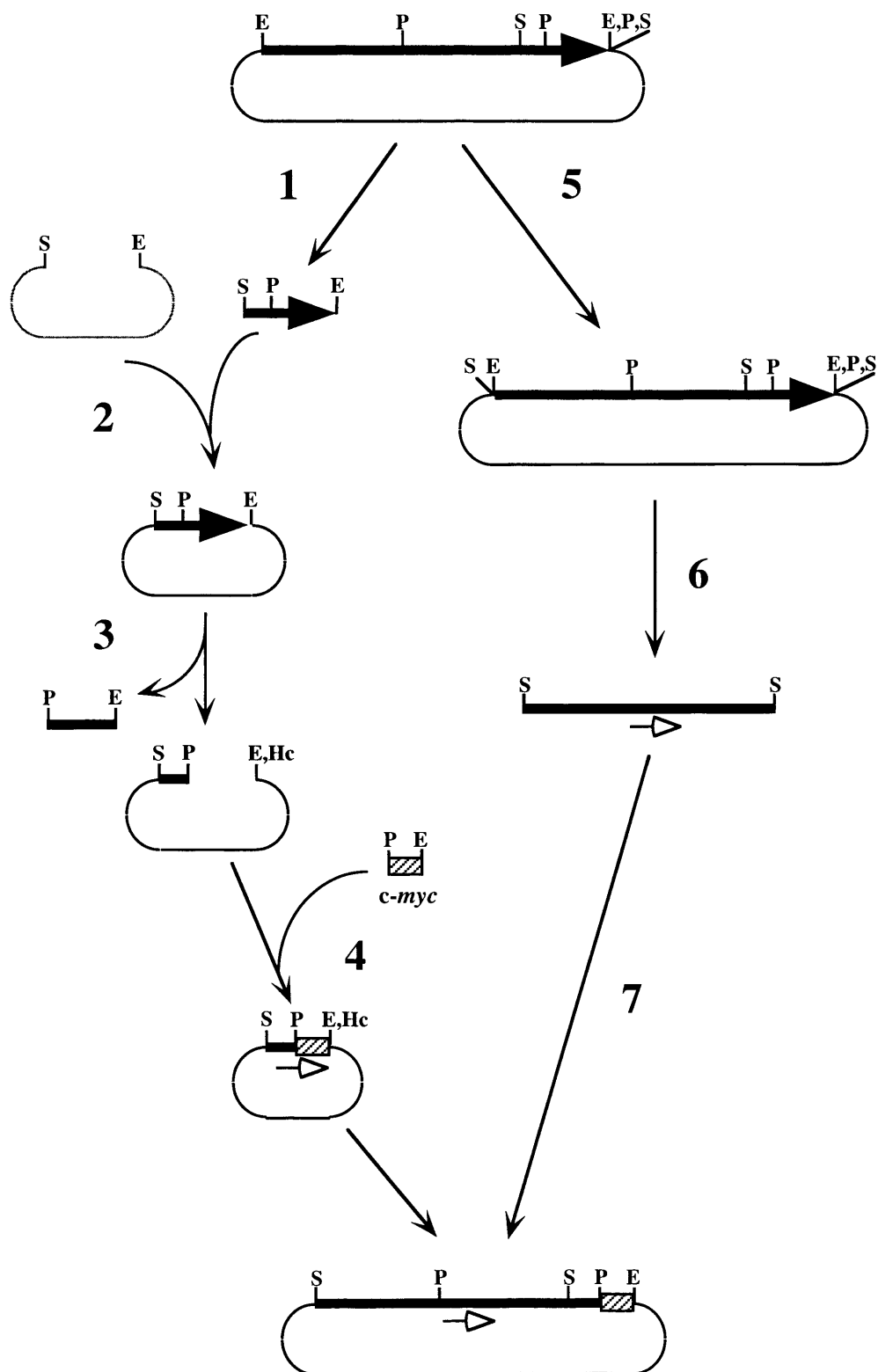


Figure 7.16 Restriction enzyme diagnostics of plasmid constructs.

Shown are electrophoretic resolutions of restriction enzyme digests of the various plasmids constructed to replace the C-termini of PKT1 and PKT2 with the c-myc epitope. Lanes M: 1 kb DNA ladder marker (GibcoBRL).

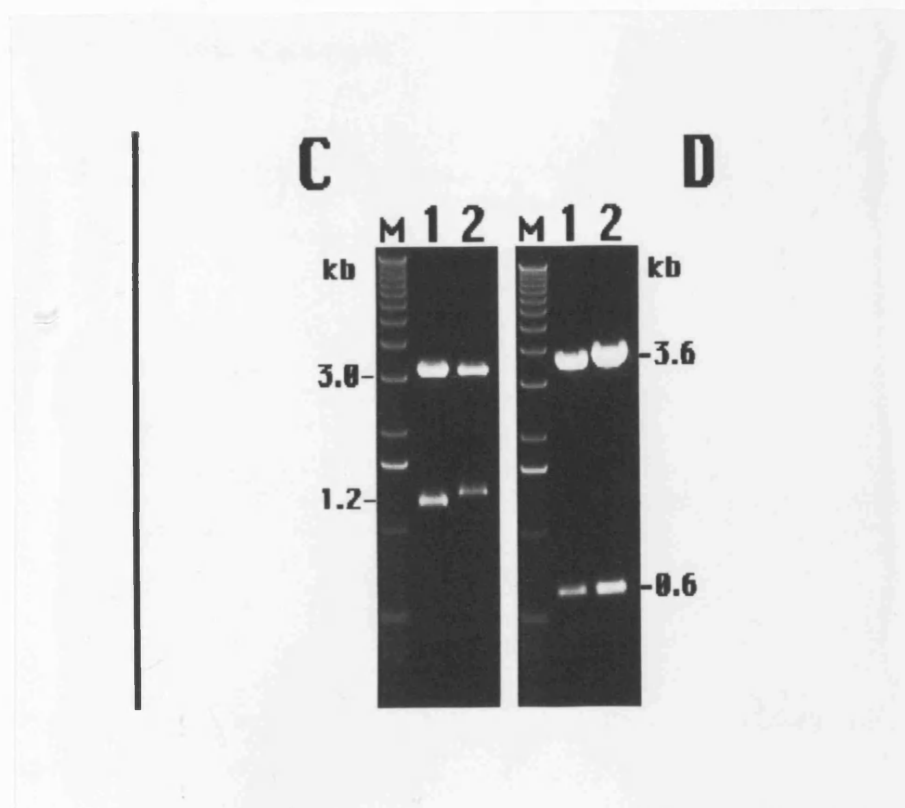
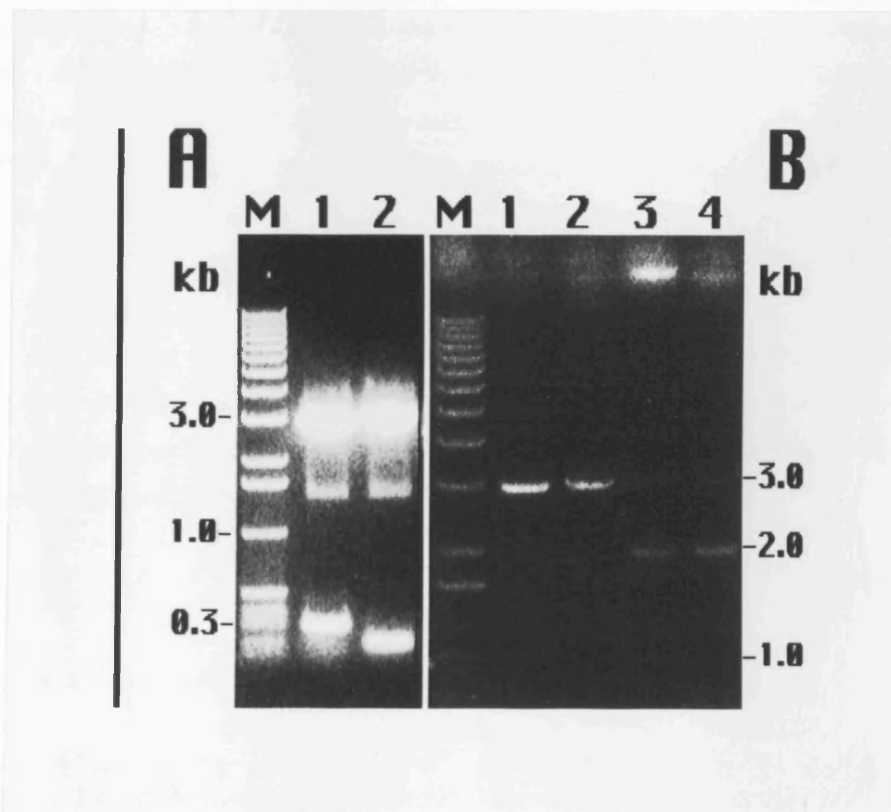
A) Digests of pSKSE-3. Lane 1: *EcoR* V and *Sac* I; Lane 2: *EcoR* V and *Pst* I.

The fainter band of fragments migrating with an estimated size of about 1.6 kb consist of undigestible denatured plasmid.

B) Digests of pSKMYC-1. Lane 1: *Pst* I plus *EcoR* V; Lane 2: *Pst* I.; Lane 3: *Xba* I; Lane 4: *BamH* I. *BamH* I and *Xba* I were not expected to cut the plasmid. The faint bands of linearized plasmid visible in these two digests should correspond to damaged molecules.

C) *Sac* I digests of pCINV1-8 (Lane 1) and pCINV1-9 (Lane 2).

D) Diagnostic of final constructs with *Pst* I. Lane 1: pC18MYC; Lane 2: pC19MYC.



For plasmids pCINV1-8 and pCINV1-9, the orientation of their inserts was simply verified by digestion with *Sac* I which was anticipated to release most of the insert as 1.2 kb and 1.3 kb fragments, respectively (Fig. 7.16C). The diagnostic digests for the constructs pC18MYC and pC19MYC used *Pst* I as it enabled their differentiation from all other plasmids. The insert derived fragments of 621 bp and the longer vector containing fragment of 3.56 kb (pC18MYC) and 3.65 kb (pC19MYC) were observed upon electrophoretic resolution of the digests (Fig. 7.16D).

This process was successful in replacing only the last 23 aminoacid-encoding codons of the cDNAs with the *c-myc* epitope (Fig. 7.17). In addition, the *c-myc* fragment was designed in order that the 11 codons encoding the epitope (Munro & Pelham, 1987) were immediately followed by a stop codon. To optimize the translation of the

Wildtype cDNA sequence

```

      G  T  G  M  G  A  A  A  V  F  E  R439          G  D  S  H457
..GGCACTGGAATGGGAGCTGCAGCTGTTTTTGAGAGG.....GGGGATAGTCATTAG..
                                                    Ter

```

c-myc tagged cDNA sequence

```

      G  T  G  M  G  A  A  E  Q  K  L  I  S  E  E  D  L  N445
..GGCACTGGAATGGGAGCTGCAGAAACAAAAGCTTATTTCTGAAGAGGATCTTAACTGAATTC.
      ACGTCTTGTTTTCGAATAAAGACTTCTCCTAGAATTGACTTAA
      (prMYC-R)                                     Ter

```

Figure 7.17 *c-myc* tagging of the 3' end of the *PKT1* and *PKT2* cDNAs.

The wildtype nucleotide sequence and corresponding translation are shown (Top) together with the modified version after insertion of the tag. The sequences derived from primers prMYC-F and prMYC-R are shown in italics. Underlined and in bold typeface are the *Eco*R I and *Pst* I sites. The termination codons are indicated by "Ter".

c-myc segment its sequence was chosen so as to include the most frequently used codons in *Arabidopsis* (Nakamura *et al.*, 1997).

Figure 7.18 Conserved physico-chemical properties in blocks A', B' and C'.

The occurrences of the blocks in 50 thiolases sequences are shown. From left to right: Block A', Block B' and Block C'. The conservation of physico-chemical properties is illustrated by the coloring. Aminoacid property groups as defined by Taylor (1986) and modified by Nicholas *et al.* (1997). Colors are: Black on Green-polar and charged residues; White on Black-hydrophobic residues; Black on Yellow-small residues; Green on Red-glycine; Green on Blue-negatively charged residues; Red on Green-amphoteric residues; Blue on Grey-aromatic residues.

1) PKT2	:	DIGIGAV	SM	ASQISDGA	TL	DIDLF	IN	A	A..	: 36
2) A.t. -PKT3	:	DIGIGAV	SM	SSQVSDGAVL	TL	DIDLF	IN	A	A..	: 36
3) B.n. -pkt	:	DIGIGAV	SM	SSQVSDGAVL	TL	DIDLF	IN	A	A..	: 36
4) C.m. -pkt	:	DIGIGAV	SM	SSQVSDGA		DIDLF	IN	A	A..	: 36
5) C.s. -pkt	:	DIGIGAV	SM	SSQVSDGAVL	TL	DIDLF	IN	A	A..	: 36
6) M.l. -pkt	:	DIGIGAV	SM	SSQVSDGA	TL	DIDLF	IN	A	A..	: 36
7) H.s. -pkt	:	DIGMACV	SM	SSQVSDGAAVTL		DIDIF	IN	A	A..	: 36
8) R.n. -pkt	:	DIGMACV	SM	SSQVSDGAAVTL		DIDIF	IN	A	A..	: 36
9) C.e. -pkt	:	NALAGAV	TL	ASQITDGA	LI	DIDVF	IN	A	A..	: 36
10) C.t. -pkcA	:	ECLAGAV	SM	ASQVSDGA	TL	DIDVF	IN	A	A..	: 36
11) S.c. -pkt	:	DIGLAV	SM	ASQVSDGAVL	TL	DIDIF	IN	A	A..	: 36
12) Y.l. -pkt	:	DIGICAV	SM	ASQISDGA	TL	DIDVF	IN	A	A..	: 36
13) B.t. -mkt	:	DVVAGAV	LM	SSFLTDGASAVL		DIDVF	IN	A	A..	: 36
14) H.s. -mkt1	:	EVVLGAV	SM	ASQVADGAVL		DMDLV	IN	A	A..	: 36
15) R.n. -mkt1	:	EVVLGAV	SM	ASQMSDGA	VLI	DMDLI	IN	A	A..	: 36
16) H.s. -mkt2	:	DVVAGAV	LM	SSFLTDGASAVL		DIDAF	IN	A	S..	: 36
17) R.n. -mkt2	:	DVVAGAV	LM	SSFLTDGASAVL		DIDAF	IN	A	S..	: 36
18) C.e. -scp	:	DVVLCAV	RM	CSPLSDGAAAVL		DVQVI	IN	C	A..	: 36
19) G.g. -scp	:	DVVLCAV	RM	CCPLSDGAAAVL		DVQVI	IN	C	S..	: 36
20) H.s. -scpx	:	ECVLAF	RM	CCPLSDGAAAVL		DIDVI	IN	C	S..	: 36
21) M.m. -scpx	:	NCVLAF	RM	CCPLSDGAC	LI	DIDVI	IN	C	S..	: 36
22) O.c. -scpx	:	ECVLAF	RM	CCPLSDGAAAVL		DIDVI	IN	C	S..	: 36
23) A.c. -kt	:	FLIAGAV	SM	ASQINDGAAVLI		QMDVI	IN	A	A..	: 36
24) A.e. -kt	:	EIVVAGAV	LM	ASQINDGAAV	VV	DIDLM	IN	A	A..	: 36
25) A.f. -kt	:	KALIVAGAV	SM	ASQVNDGAAVVL		DIGLI	IN	A	A..	: 36
26) A.s. -kt	:	EIVVAGAV	SM	ASQINDGAAV	VV	DVDLF	IN	A	A..	: 36
27) A.sp-kt	:	DIVVAGAV	SM	ASQINDGAAVLM		QVDLI	IN	A	A..	: 36
28) E.c. -kt	:	QACIVGAV	HM	SSALSDGAAVLI		DIGVF	IN	A	A..	: 36
29) M.t. -kt	:	CAFISA	TV	ACPLNDGAAVLI		DIDLV	IN	A	A..	: 36
30) M.t. -kt2	:	DVLEAV	SM	SSQITDGA	AVL	DIDLF	IN	A	A..	: 36
31) P.f. -kt	:	DVFWGAV	HM	SSQITDGA	AVL	DIDFI	IN	A	A..	: 36
32) P.p. -kt	:	FLIAGAV	SM	ASQVNDGAAV	LI	DFDVI	IN	A	A..	: 36
33) C.t. -pat	:	EIVVAGAV	SM	ASQINDGAAVLI		RDFFL	IN	A	S..	: 36
34) H.s. -mat	:	DVMVAGAV	SM	ASTLNDGAAVLI		DIAMV	IN	A	S..	: 36
35) R.n. -mat	:	SIVVAGAV	LM	ASQINDGAAVAVL		QVDIF	IN	A	A..	: 36
36) H.s. -cat	:	DVVVAGAV	SM	ASSISDGA	AVL	QVDYF	IN	A	A..	: 36
37) R.s. -cat	:	DVVVAGAV	SM	ASQINDGAAVLI		QVDYF	IN	A	S..	: 36
38) S.c. -cat	:	EIVVAGAV	SM	ASTLNDGASALVL		QVDYF	IN	A	S..	: 36
39) S.p. -cat	:	DIVVAGAV	SM	APQVNDGAFVL		DIDLF	IN	A	A..	: 36
40) B.s. -at	:	DVILAGAV	SM	ASQINDCAVLI		DIDLI	IN	A	A..	: 36
41) C.a. -at	:	NALTAGAV	TL	ASQITDGA	AVL	DVDVF	IN	A	A..	: 36
42) C.e. -at	:	QSTVAGAV	LM	ASQINDGA	AVL	DIDLI	IN	A	A..	: 36
43) E.c. -at	:	DIVVAGAV	LM	ASQINDGA	AVL	DIDLI	IN	A	A..	: 36
44) H.i. -at	:	RIVVAGAV	SM	ASQINDGA	AVL	DIDLV	IN	A	A..	: 36
45) P.d. -at	:	KIVVAGAV	SM	ASQINDGA	AVL	DIDLV	IN	A	A..	: 36
46) R.m. -at	:	DILLAGAV	SM	CSQITDGA	AVL	DFDLF	IN	A	S..	: 36
47) S.s. -at	:	DIVVAGAV	LM	ASQINDGA	AVL	DIDLI	IN	A	A..	: 36
48) T.t. -at	:	EIVLAGAV	SM	ASQINDGA	AVL	DIDLV	IN	A	A..	: 36
49) T.v. -at	:	SIVLAGAV	SM	ASQINDGA	AVL	DIDLV	IN	A	A..	: 36
50) Z.r. -at	:	SIVLAGAV	SM	ASQINDGA	AVL	DIDLV	IN	A	A..	: 36

G E

F

7.9 Development of a new and more powerful signature for thiolases

The three blocks of conserved sequence A, B and C in PKT2 (Section 7.5) were examined for the properties of their constituent residues. It was observed that in each there was a very high degree of conservation of physico-chemical properties of the residues, as defined by Taylor (Taylor, 1986), rather than identity, at most positions. Indeed, the blocks were re-defined according to conservation of these properties. Of the resulting blocks, designated A', B' and C', the latter two had some alterations relative to the limits of blocks B and C. An extended list of the re-defined block sequences found in thiolases is shown in Figure 7.18. The physico-chemical properties of the aminoacids found at each block are summarized in Figure 7.19.

The level of conservation of the blocks A'-C' was comparable to those of the PROSITE signatures. The signatures 1 and 3 are known to generate a number of false hits when used to scan the protein databases Swiss-Prot and TrEMBL (Bairoch & Apweiler, 1998). Also, signature 3 is absent from some thiolase sequences including most

Block A'

Polar/Charged-X-Hydroph.-Small-Hydroph.-G-X-E-X-Hydroph.

Block B'

2(Small)-X-Hydroph.-Small-Small/Amphot.-3(Small)-4(Hydroph.)

Block C'

Polar/Charged-Hydroph.-X-2(Hydroph.)-E/D.-X-Amphot.-E/D.-Small-F

Figure 7.19 Conserved physico-chemical properties of the residues in blocks A', B' and C'.

The aminoacid property groups as defined by Taylor (Taylor, 1986) and modified by (Nicholas *et al.*, 1997) were used. These included: Polar/ Charged (YWHKREQDNSTBZ); Hydrophobic (ILVCAGMFYWHTP); Small (VCA-GDNSTP); Amphoteric (HREQDNBZ); Abbreviations are: Hydroph.-hydrophobic; Amphot.-amphoteric.

sterol carrier proteins X (cf. human and mouse SCP-X, Fig. 7.5). For these reasons the development of a new signature sequence combining high accuracy with a greater degree of comprehensiveness would be useful. The three conserved blocks A'-C' defined candidates for new signatures and this possibility was investigated. The process of selection included several phases. Initially, the Swiss-Prot and TrEMBL databases (Bairoch & Apweiler, 1998) were

scanned with minimal consensus derived from the blocks as defined by the alignment in Figure 7.5. The results are shown in Table 7.4. To improve the specificity and inclusiveness towards thiolases, the minimal consensus were then subjected to systematic alterations and thorough testing of variants by scanning the databases. The alterations tested included changes at all conserved positions as well as additions and deletions. Final versions of the consensus were eventually obtained characterized by high specificity and comprehensiveness of detection (Table 7.4).

The final consensus C generated a lower number of hits but displayed maximum specificity with 100% of the detected sequences corresponding to known or putative thiolases (Table 7.4). Both the consensus B and A were clearly superior to the PROSITE signature 3. The performance of consensus A was comparable to that of signature 1 but failed to detect sterol carrier proteins. The final consensus B was found to be better than all three PROSITE signatures in effectiveness of detection (81 hits) and nearly as selective (94% accuracy) as the best of the PROSITE signature sequences (97% accuracy). Importantly, it is capable of detecting all types of thiolases including the sterol carrier proteins. Of the five false hits it generated, three concerned proteins involved in lipid transport and metabolism which share a number of structural features with thiolases (data not shown).

7.10 Summary

The wildtype transcripts mP18 and mP19 mapping to the genomic region tagged in AtEN101 were analyzed to identify the proteins that they potentially encode. One strong candidate ORF was found for each transcript based on several criteria including their positioning and extension relative to the corresponding RNA molecules, their codon composition, the sequences containing their proposed translation initiation codon (Section 7.2), and their strong similarity to known proteins (Section 7.3). The 414 residues-long ORF derived from mP18, PKT1, is essentially a truncated form of PKT2 (from mP19) missing the N-terminal 43 aminoacid residues. PKT1 and PKT2 were found to be highly similar to thiolases throughout their length (Section 7.4), sharing with these enzymes a number of well conserved primary structure features including the three catalitically most important aminoacid residues (Section 7.5).

Several thiolase types exist including acetoacyl-CoA thiolase (EC 2.3.1.9) normally involved in biosynthetic pathways, and 3-ketoacyl-CoA thiolase (EC 2.3.1.16), one of the enzymes of the β -oxidation cycle. In addition, in eukaryotes these enzymes are distributed in three cellular compartments. To confirm the potential enzymatic function of the ORFs, predictions of their secondary structure were obtained by several methods and compared to the known secondary structure of the yeast peroxisomal 3-ketoacyl-CoA thiolase (Section

	Total	Thiolase	Other	Rel.
Consensus A				
[KSEDQN]-[SIAVLC]-[IVAMFGC]- [IVLM]-[AVCG]-[GALC]-G-X-E-X-[ML]	89	65 (73 %)	24	?
[IVL]-X(5)-[DQKSERH]-[MIVSLAC]-X- [IVLM]-X-[GACLVIM]-G-X-E- [NSHLTRK]-M	75	69 (92 %)	6	3
Consensus B				
[CSAT]-X-[TLIVM]-X-[DN]-G-[SAVG]- [CASG]-[CAVG]-[MAVIL]-[MAVIL]- [MAVIL]	91	78 (86 %)	13	6
[ICSA]-[PCSAT]-X-[TLIVM]-X-D-G- [SAVG]-[CASG]-[AGCVTMF]- [FTAMVLI]-[MVILA]-[MAVLI]	86	81 (94 %)	5	3
Consensus C				
DRSQYE]-[IVMFL]-[DEAG]- [FVIMYL]-[SFVIMYLW]-[DE]-X-[NH]-E- A-F-[SA]	63	63 (100 %)	0	0
[WHKNTDRSQYE]-[AGCYWHIVMFL]- [DEAG]-[FVIMYL]-[SFVIMYLW]-[DE]- X-[NH]-E-A-F-[SAR]	63	63 (100 %)	0	0
PROSITE Signature 1				
[LIVM]-[NST]-X(2)-C-[SAGLI]-[ST]- [SAG]-[LIVMFYNS]-X-[STAG]-[LIVM]- X(6)-[LIVM]	83	75 (90 %)	8	2
PROSITE Signature 2				
N-X(2)-G-G-X-[LIVM]-[SA]-X-G-H-P-X- G-X-[ST]-G	76	74 (97 %)	2	1
PROSITE Signature 3				
[AG]-[LIVMA]-[STAGLIVM]-[STAG]- [LIVMA]-C-X-[AG]-X-[AG]-X-[AG]-X- [SAG]	72	59 (82 %)	13	0

Table 7.4 Efficiency of detection of thiolases by signature sequences.

Results obtained with the initial (top lines) and refined version (lower lines) of the consensus sequences for blocks A, B and C. The efficiency is indicated by the total number of detected proteins (Total), and the number (and percentage) of thiolases in the detected proteins (Thiolase). Also indicated are the number of non-thiolase proteins detected (Other), and of these, those that are functionally related to thiolases (Rel.). Similar values were calculated for the PROSITE signatures 1, 2 and 3.

7.6). An overall matching of the elements was observed. The degree of correspondence was even higher relative to similar predictions performed for the yeast enzyme. Taking advantage of the existence of a crystallographic description of the same yeast thiolase, a 3D model was constructed for PKT1/PKT2 by homology-threading (Section 7.7). The resulting model was highly similar to the structure of the yeast enzyme, most differences being located in its periphery, away from the pocket containing the active site which was well conserved. Together, the data were highly indicative that the two putative proteins should have catalytic activities similar to that of the yeast peroxisomal thiolase.

The possible intracellular site of activity of the putative proteins was analyzed in Section 7.8. The higher degree of similarity to peroxisomal 3-ketoacyl-CoA thiolases rather than to other thiolase types indicated that PKT1 and PKT2 should also be active in the peroxisome. In accordance with this hypothesis, a match to the peroxisomal targeting signal PTS2 was present in the N-terminal region of PKT2. However, PKT1 lacked any recognizable peroxisomal targeting signal. This made it of interest to experimentally determine the cellular localization of PKT1. A suitable strategy to address this problem would be the expression in transformed plants of a structurally modified PKT1 protein, easily distinguishable from PKT2. Within this context, the cDNAs in clones pCEN1-8 and pCEN1-9 were modified by replacement of the segments encoding the C-terminal 23 aminoacid residues of the PKT proteins with a short DNA fragment encoding the *c-myc* epitope (Munro & Pelham, 1987). Future work involving transformation of plants with the modified cDNAs under the control of strong promoter should enable the intracellular localization of the modified PKT1 and PKT2 proteins by immunochemical methods.

Finally, three segments highly conserved in thiolases that had been identified during the comparative analysis of PKT1/PKT2 were assessed for their potential development as signatures for these enzymes (Section 7.9). This work was justified as it became clear that the currently recognized PROSITE signatures for thiolase enzymes (Bairoch & Bucher, 1994) had limitations of accuracy and/or efficiency of detection. It was possible to develop from one of the segments, a new signature sequence, consensus B, with the highest recognition capacity for thiolases and very high accuracy (94%).

Chapter 8

The promoter regions of *PKT1* and *PKT2*

8.1 Introduction

The identification of the regulatory regions controlling the expression of the reporter gene in AtEN101 was an important objective of the present work (cf. Section 1.4.2). This chapter is concerned with the structural study of the regions containing the putative promoters of *PKT1* and *PKT2* and the testing of their potential regulatory activities of transcription. A further objective of the chapter is the investigation conducted on the possible functional relation of the promoters of *PKT1* and *PKT2* and the expression of *gus* in AtEN101.

The 5' RACE of fusion transcripts (Section 6.2) and the study of the T-DNA structure (Section 6.3) had established a likelihood of a link between the activity of the promoter of *PKT1* and the expression of *gus*. However, the furthest 5' limit of the fusion transcripts defined by RACE and coinciding with the 5' end of mP18 (Section 6.2.4) was located considerably downstream from what was anticipated based on the Northern hybridization analysis by Topping *et al.* (1994). Also, the characterization of the *PKT1* transcripts reported in Chapters 5 and 7 was insufficient to determine the transcription start site of the gene. This raised the possibility of the existence of additional exons for *PKT* genes, and/or regulatory elements further upstream that might contribute for the structure of (the) fusion transcripts. Detailed characterization of the region upstream from the 5' end of the mP18 transcripts should provide information with regard to such potential exons and/or regulatory *cis*-elements. With this objective, about 1.2 kb of the genome upstream from the 5' end of mP18 were sequenced (Section 8.2). The results of the sequence analysis prompted a reassessment of the size of the fusion transcripts in AtEN101 which was performed by Northern hybridization (Section 8.3).

As for *PKT1*, the transcription start site of *PKT2* was unknown, hence there was uncertainty with regard to the location of its promoter region. Nevertheless, the arrangement of the two genes indicated that the promoter should overlap the first intron of *PKT1*. As a test of this possible organization of *PKT2* the presence of corresponding transcripts in AtEN101 was investigated by RT-PCR (Section 8.4.1). For the same reason, the structure of the segment comprising the first exon and intron of *PKT1* up to the 5' end of mP19 was searched for matches to known *cis*-acting regulatory elements of transcription that might be indicative of the presence of the *PKT2* promoter (Section 8.4.2). Several such matches were indeed found.

The potential of the region upstream of mP19 and of segments upstream from mP18 to drive transcription *in vivo* was investigated in transient expression assays (Section 8.5). The possible presence of a cryptic promoter in the genomic segment E1.4E1.5 that might drive the expression of the *gus* gene located at the T-DNA right junction in AtEN101 was similarly tested. For these studies plasmid constructs were assembled consisting of the relevant genomic segments fused to a downstream promoterless reporter gene *gus*. The constructs were delivered by biolistics to living plant tissue and subsequent expression of GUS was monitored.

8.2 Sequencing the putative promoter region of *PKT1*

8.2.1 Generation of Tn1000 insertion-derivatives of pENB8

The region upstream from the E5.2 site which should contain the promoter of *PKT1* was known to be very poor in restriction enzyme sites (Section 5.3.3). Sequencing this region based on a directed subcloning strategy would be difficult. For this reason another cost-effective alternative was chosen based on insertion mutagenesis. The strategy consisted of generating a population of pENB8-derived plasmids randomly mutagenized by insertion of a variant of transposon Tn1000. Sequencing could then proceed from the common terminal regions of the transposon in a selection of the plasmids using a same set of primers.

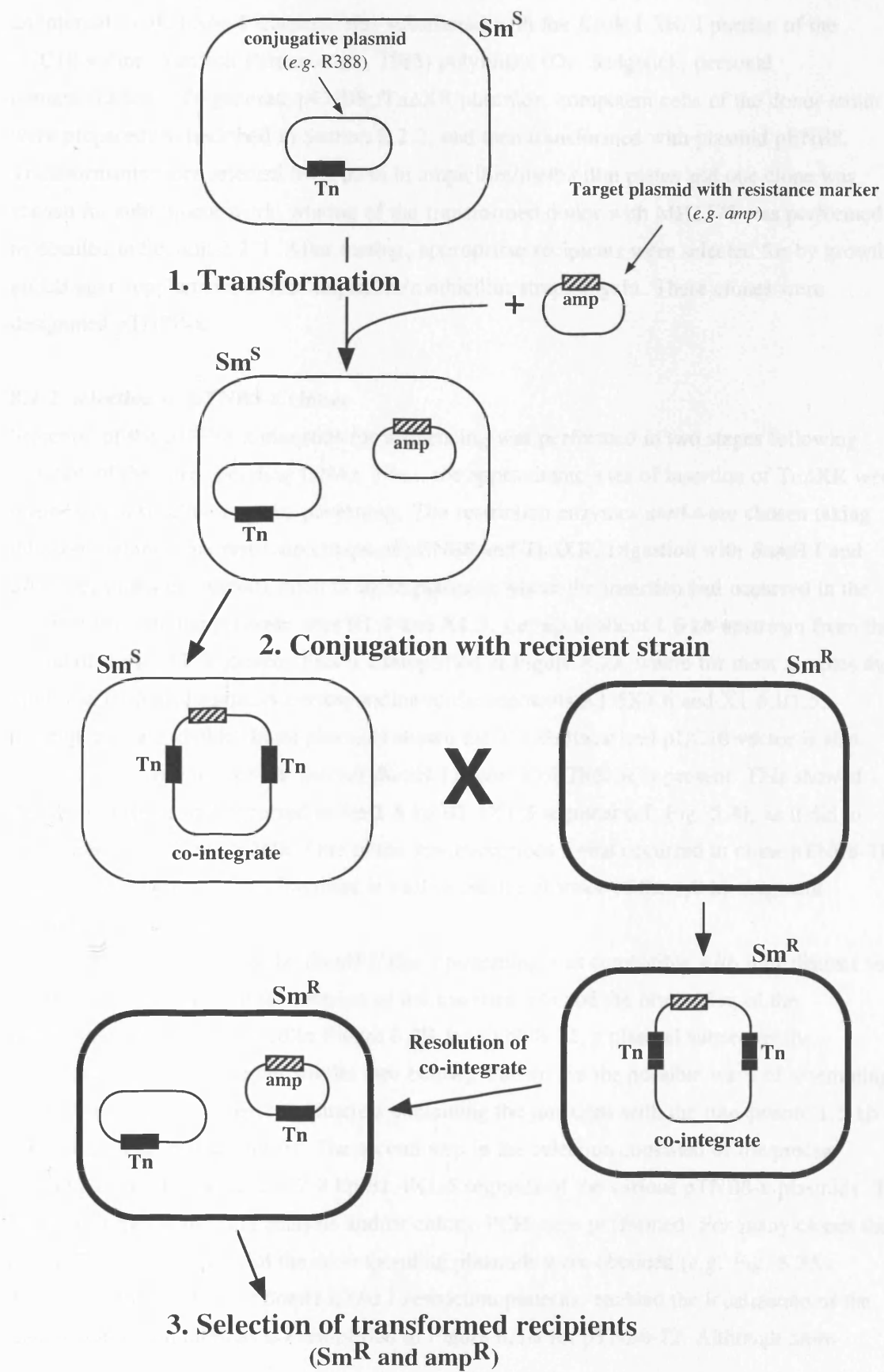
The 5980 bp-long Tn1000 is a naturally occurring transposon of *E. coli* harboured in its F sex factor (Guyer, 1978). It can be easily transposed from a conjugative plasmid to another plasmid, the insertions being quasi random (Sedgwick & Morgan, 1994). The transposition involves the formation of a co-integrate, a large plasmid consisting of the host sex factor and the target plasmid sequences joined by duplicated copies of the transposon (Fig. 8.1). Resolution of the co-integrate regenerates the conjugative plasmid and results in a derivative of the target plasmid carrying a copy of the transposon. These properties were exploited by Sedgwick *et al.* (1994) to develop a system for the generation of populations of transposed target plasmids. The system is based on a donor strain of *E. coli* hosting the conjugative plasmid R388 and an associated Tn1000-type transposon, and on a recipient strain resistant to a selection marker (Fig. 8.1). The first step in the procedure consists of the transformation of the donor bacterial strain with the target plasmid. Mating of the transformed donor and recipient strains results in the transfer of co-integrates when transposition is simultaneously taking place in the donor cells. Resolution of the co-integrates in the recipients regenerates the R388 factor and releases a target plasmid modified by the insertion of one copy of the transposon. With appropriate selection, the donor cells and those recipients not harbouring target plasmids are eliminated enabling the specific isolation of the desired clones.

The Tn1000 variant utilized, TnΔXR, hosted in the donor strain DH1, and a recipient strain, MH1578, were kindly supplied by Dr. Sedgwick (London). MH1578 is a streptomycin

Figure 8.1 Strategy for generation of a population of plasmids mutagenized with Tn Δ XR.

This strategy, devised by Sedgwick and Morgan (1994), consists of:

- 1) transformation with the plasmid to be mutagenized of a resistance marker-sensitive (*e.g.* streptomycin-Sm) strain harbouring Tn Δ XR in a conjugative plasmid. The transforming plasmid should include a second resistance marker (*e.g.* ampicillin, amp). Tn Δ XR will transpose into the plasmid of interest through the formation of a co-integrate. This co-integrate consists of a copy of each of the two types of plasmid held together at their junctions by two copies of the transposon.
- 2) If simultaneous conjugation of the plasmid host bacteria to a recipient occurs, the co-integrate will be transferred to the latter where it will be resolved. Resolution of the co-integrate restores the original conjugative plasmid and generates a target plasmid mutagenized by insertion of a copy of Tn Δ XR.
- 3) By using a recipient strain resistant to the marker to which the donor strain is sensitive (*e.g.* Sm), it is possible to selectively eliminate the donor cells and recipient cells not harbouring a suitably mutagenized target plasmid by growth of the conjugation mix in an appropriately supplemented medium.



resistant version of DH1, and Tn Δ XR is a 4888 bp replacement derivative of Tn1000 where an internal *EcoR* I-*Xba* I fragment was substituted with the *EcoR* I-*Xba* I portion of the pUC18 vector (Yanisch-Perron *et al.*, 1985) polylinker (Dr. Sedgwick, personal communication). To generate pENB8::Tn Δ XR plasmids, competent cells of the donor strain were prepared as described in Section 2.2.2, and then transformed with plasmid pENB8. Transformants were selected by growth in ampicillin/methicillin plates and one clone was chosen for subsequent work. Mating of the transformed donor with MH1578 was performed as detailed in Section 2.2.3. After mating, appropriate recipients were selected for by growth on LB agar supplemented with ampicillin/methicillin/streptomycin. These clones were designated pTNB8-x.

8.2.2 Selection of pTNB8-x clones

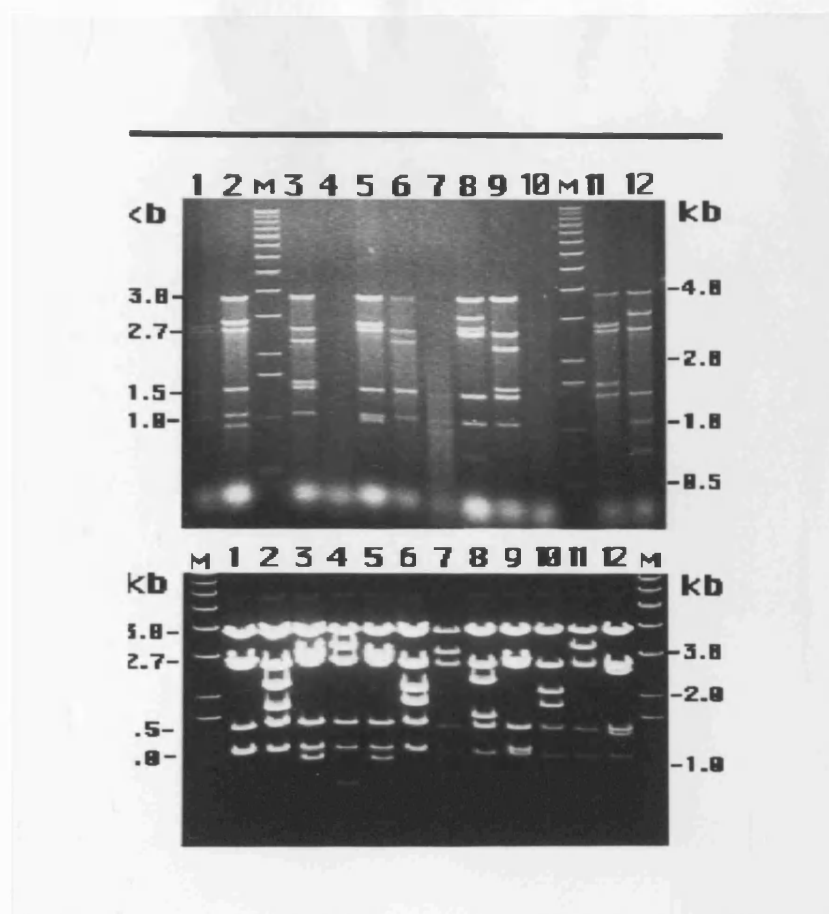
Selection of the pTNB8-x plasmids for sequencing was performed in two stages following isolation of the corresponding DNAs. First, the approximate sites of insertion of Tn Δ XR were mapped by restriction enzyme patterning. The restriction enzymes used were chosen taking into consideration the restriction maps of pENB8 and Tn Δ XR. Digestion with *Bam*H I and *Xba* I permitted the identification of those plasmids where the insertion had occurred in the segment between the genomic sites B1.4 and X1.5, *i.e.* up to about 1.6 kb upstream from the 5' end of the mP18 sequence. This is exemplified in Figure 8.2A where for most samples the 1.0 kb and 1.5 kb fragments corresponding to the segments X1.5X1.6 and X1.6.B1.5, respectively, are visible. In all plasmids shown the 2.7 kb linearized pUC18 vector is also intact (2.7 kb) and the 3.8 kb internal *Bam*H I segment of Tn Δ XR is present. This showed that the insertions had occurred in the 2.8 kb B1.4X1.5 segment (cf. Fig. 5.8), as it did in most clones surveyed (>100). One of the few exceptions found occurred in clone pTNB8-71 (Fig. 8.2A) where a 2.8 kb fragment is visible and the absence of the 1.0 kb fragment (X1.5X1.6) is noticeable.

For most plasmids the *Bam*H I/*Xba* I patterning was compatible with four distinct sites of insertion depending on the location of the insertion site and the orientation of the transposon. This is illustrated in Figure 8.2B for pTNB8-72, a plasmid subsequently sequenced and used for other studies (see below). Shown are the possible ways of assembling the two restriction fragments of interest containing the junctions with the transposon (1.5 kb and 2.4 kb; Fig. 8.2A, lane 6). The second step in the selection consisted of the precise mapping of the Tn Δ XR in the 2.8 kb B1.4X1.5 segment of the various pTNB8-x plasmids. To this end further restriction analysis and/or colony-PCR were performed. For many clones the *Sma* I restriction profiles of the corresponding plasmids were obtained (*e.g.* Fig. 8.3A). These, together with their *Bam*H I/*Xba* I restriction patterns, enabled the localization of the transposon within them as is exemplified in Figure 8.3B for pTNB8-72. Although more

Figure 8.2 Mapping Tn Δ XR insertion sites in pTNB8-x plasmids by digestion with *Bam*H I/*Xba* I.

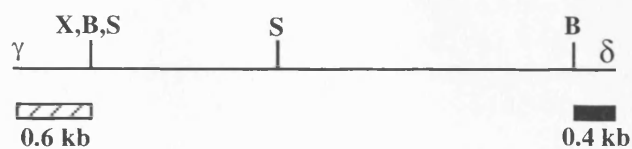
- A)** Examples of *Bam*H I/*Xba* I digests of TnB8-x plasmids resolved in agarose gels. The sizes of the vector-derived fragments (2.7 kb), the internal Tn Δ XR *Bam*H I fragment (3.8 kb), and of the smaller internal fragments (non-mutagenized) of the plasmid inserts (1.0 kb and 1.5 kb) are indicated. The other two fragments visible are composites of the terminal regions of the transposon and the adjacent segments of the 2.8 kb *Bam*H I/*Xba* I fragment. The only exception is visible in lane 11 (Top gel) where Tn Δ XR was inserted in the 1.0 kb fragment.
- B) Top:** Restriction map of Tn Δ XR (Sedgwick & Morgan, 1994) where the γ and δ termini are identified. The terminal 0.6 kb and 0.4 kb segments contributing to the junction fragments in the digests are represented by the hatched and filled boxes, respectively. Restriction enzyme sites are labelled: B-*Bam*H I, S-*Sma* I, X-*Xba* I.
- Bottom:** The four possible interpretations of the *Bam*H I/*Xba* I restriction enzyme pattern of pTNB8-72 (see top gel above, lane 6) are shown. The digest of the plasmid generated two transposon junction fragments of 1.5 kb (distinct from the 1.5 kb X1.6B1.5 fragment) and 2.4 kb. For the junction fragments, the Tn Δ XR 0.4 kb and 0.6 kb segments are represented by the boxes, and the sizes of the segments derived from the 2.8 kb *Bam*H I/*Xba* I fragment are indicated. The internal 3.8 kb *Bam*H I fragment of the transposon is omitted as indicated by the periods. The pBluescript sequences are shown as a black bar.

A)



B)

Tn Δ XR



Possible modes of insertion in pTNB8-72

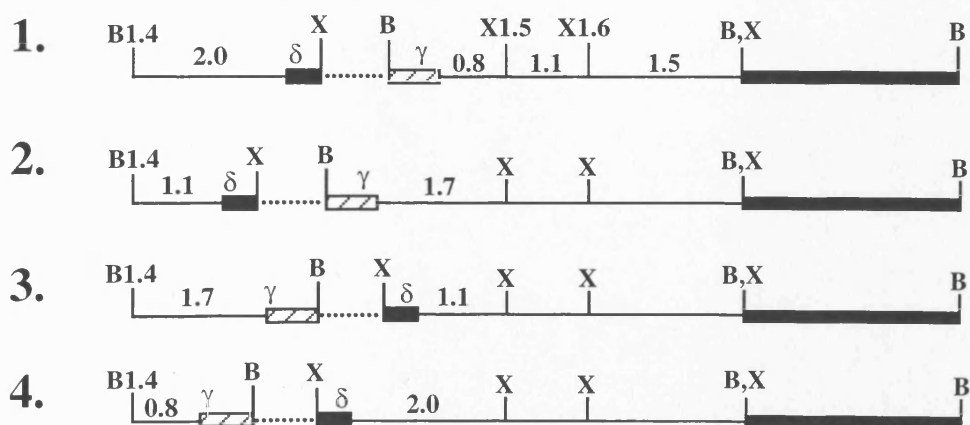
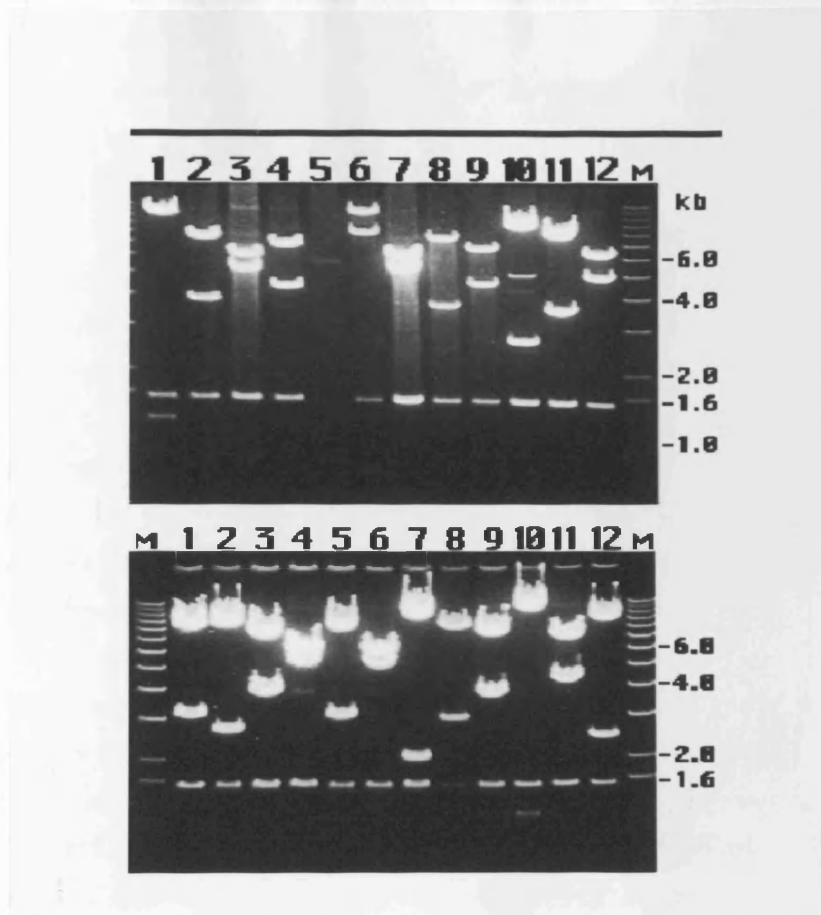


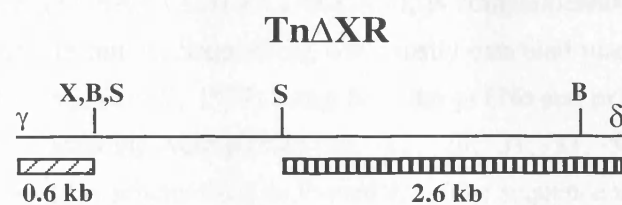
Figure 8.3 Mapping Tn Δ XR insertion sites in pTNB8-x plasmids by digestion with *Sma* I.

- A)** Examples of gels of *Sma* I digested pTNB8-x plasmids. The size of the 1.6 kb internal *Sma* I fragment of Tn Δ XR is indicated. Together with the information derived from the *Bam*H I/*Xba* I digests (*e.g.* Figure 8.2), the *Sma* I restriction patterns permitted the determination of the site of transposon insertion and its orientation.
- B)** Exemplification of the mapping of the transposon in a pTNB8-x plasmid. The observed sizes of the junction fragments of pTNB8-72 were 4.6 kb and 6.9 kb (see lane 9, top gel above). The anticipated *Sma* I patterns for each of the models of insertion derived from pTNB8-72 by digestion with *Bam*H I/*Xba* I (cf. Figure 8.2B) are shown and the sizes of the expected fragments are indicated. Clearly, only model 1 matches the observed pattern.

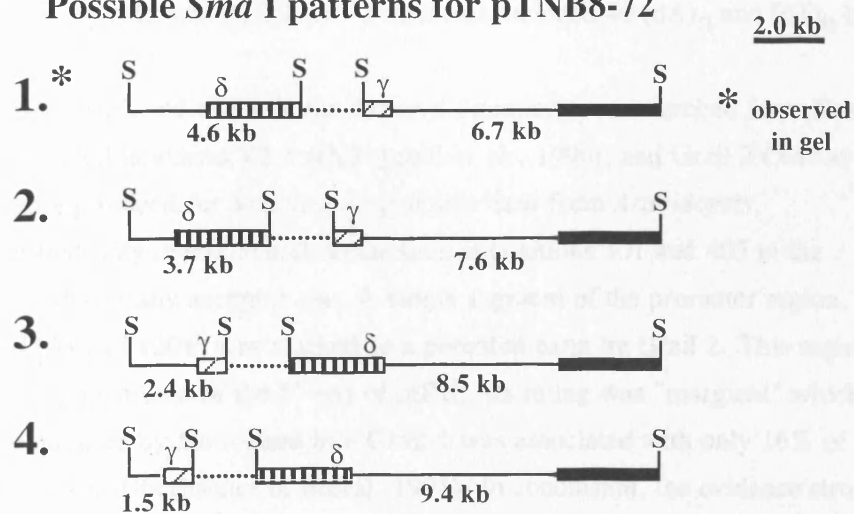
A)



B)



Possible *Sma* I patterns for pTNB8-72



troublesome than colony-PCR, this method was highly reliable and usually indicative of the orientation of the transposon.

Because of its higher precision and ease, colony-PCR was the tool used to map the location of the transposon in most clones. Tn1000 and derived transposons feature at both ends a 35 bp-long terminal repeat (AACGTACGTTTTCGTTCCATTGGCCCTCAA-ACCCC). For colony-PCR a 33 residues-long primer, prTN1000-RPT, was designed consisting of the last 25 residues Tn Δ XR terminal repeat preceded at its 5' end by an unrelated 8 bp segment including a *Bam*H I site (see sequence in Appendix 5). For amplification of the segment between the *Bam*H I site B1.4 and the transposon, the primer prTN1000-RPT was used in conjunction with the universal reverse primer, as described in Section 2.5.5.4. The products of the amplification reactions were sized by agarose gel electrophoresis (*e.g.* Fig. 8.4A). After the initial mapping, a more precise ordering of the sites of Tn Δ XR insertions was achieved by electrophoresis of aliquots of the PCR reactions loaded in order of the apparent size of their products (*e.g.* Fig. 8.4B).

8.2.3 The sequence of the *PKT1* promoter region

A set of clones where the single transposons were regularly spaced along the length of the B1.4X1.5 segment were selected for sequencing. Two primers were designed from the two terminal regions of Tn Δ XR. One of the primers, prTN δ (AGGGGAACTGAGAGCTCTA), corresponds to a sequence located 66 bp upstream from the δ end of the transposon. The second primer, prTN γ (TCAATAAGTTATACCAT), is complementary to a sequence 36 bp downstream from the γ terminus. Sequencing was mostly executed manually by the dideoxy-termination method (Sanger *et al.*, 1977) using both the prTN δ and prTN γ primers. The plasmids chosen for sequencing were pTNB8-24, -26, -28, -31, -33, -54, -64, -66, and -72, and the strategy followed is schematized in Figure 8.5. The sequence up to 1240 bp upstream from the 5' end of the mP18 was determined on both strands (Fig. 8.6). Another 567 bp further upstream were also sequenced but, for the most part, from a single strand (data not shown). The 1240 bp segment was 71.3% A+T-rich and included 40 (dA)_n and (dT)_n tracts ($n \geq 4$).

Potential splice sites and exons in the forward direction were searched for with two neural network tools, NetPlantGene V2.1 (Hebsgaard *et al.*, 1996), and Grail 2 (Xue *et al.*, 1994). Both tools are optimized for searches on genomic data from *Arabidopsis*. NetPlantGene identified only two potential donor sites at positions 101 and 405 in the sequence. It failed to detect any acceptor site. A single segment of the promoter region, between positions 1104 and 1201, was marked as a potential exon by Grail 2. This segment was located only 40 bp upstream of the 5' end of mP18. Its rating was "marginal" which on a test set of human sequences by the related tool Grail 1 was associated with only 16% of correctly predicted exons (Uberbacher & Mural, 1991). In conclusion, the evidence strongly

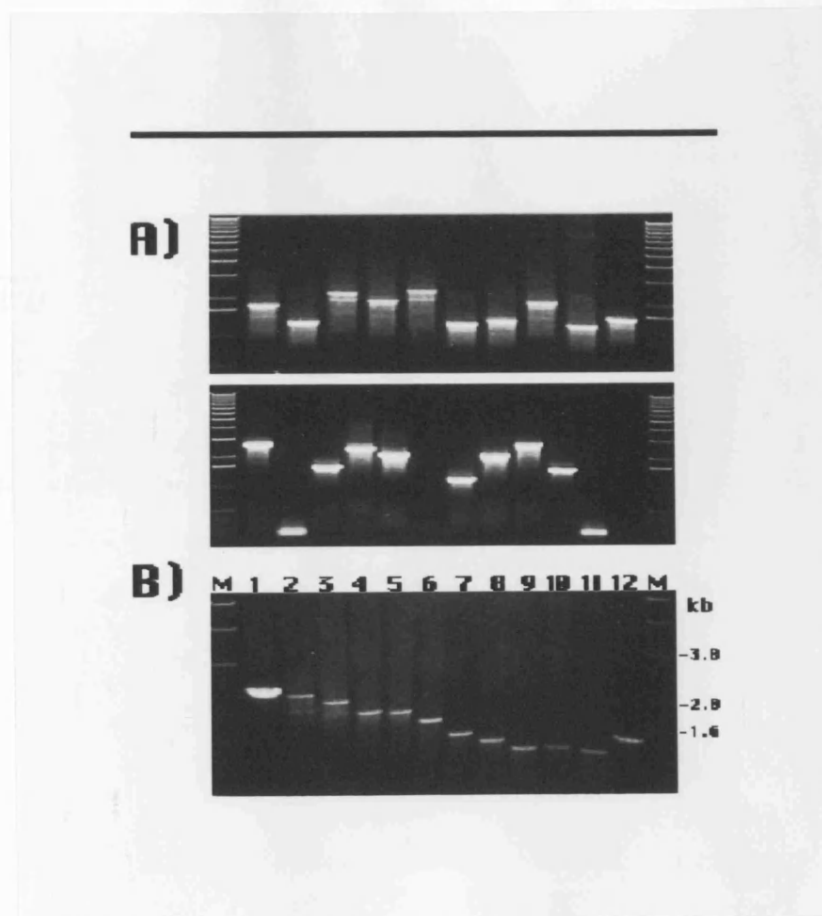


Figure 8.4 Mapping the location of Tn Δ XR inserts in pTNB8x plasmids by colony-PCR.

A) Colony-PCR was performed on *E. coli* clones of pTNB8-x plasmids. The universal reverse primer and the prTN1000-RPT were used for the amplification. Shown are examples of two of the gels used to resolve the products of amplification.

B) Example of a gel utilized for fine mapping of the transposon insertion sites. To more accurately determine the relative positioning of the transposon sites in the pENB8 insert, products of amplifications were loaded in the gels in order of their estimated distance to the B1.4 site. The amplification products obtained from several plasmids used for sequencing can be seen in the gel shown. Lane 2: pTNB8-64; Lane 5: pTNB8-54; Lane 6: pTNB8-28; Lane 7: pTNB8-33; Lane 9: pTNB8-72; Lane 10: pTNB8-26; Lane 12: pTNB8-78.

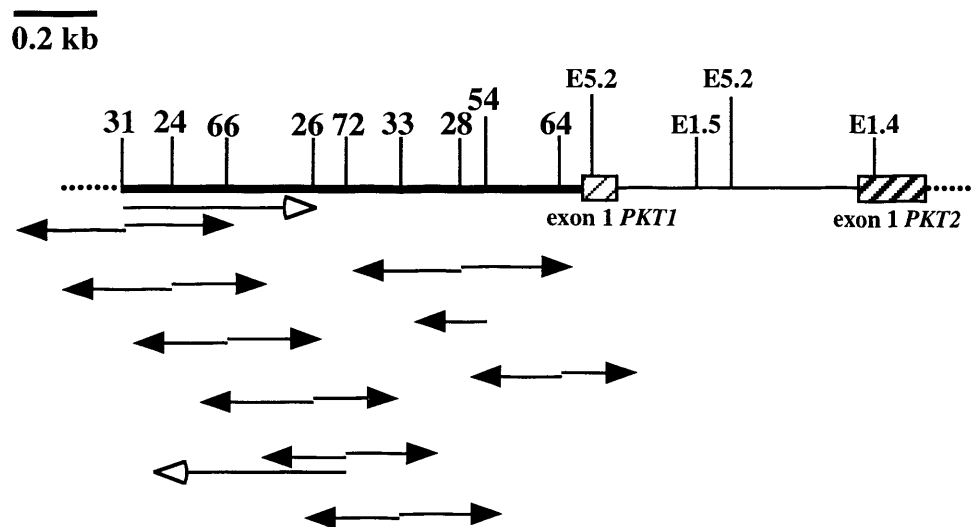


Figure 8.5 Sequencing strategy for the promoter region of *PKT*.

Sequencing was performed on the transposon-mutagenized derivatives of pENB8 and a composite sequence assembled. The arrows indicate the length and direction of sequencing. Filled and open arrowheads denote the mode of sequencing, manual and automatic respectively. The composite sequence is shown as a thick continuous line. The sites of Tn Δ XR insertion in the various pTNB8-x plasmids are identified by the serial number of the clones. The first exons of mP18 and mP19 are also shown as boxes. Other sequences are shown as thin lines.

indicated that no exons of *PKT1* and/or *PKT2* existed upstream from those previously identified in mP18 and mP19.

Both the experimental data and the sequence analysis indicated that the first exon of mP18 coincided with the first exon of *PKT1*. For this reason the sequence upstream from mP18, which should contain the promoter of *PKT1*, was examined for matches to known transcription factor binding sites and other known *cis*-acting sequences of vascular plants. Several on-line search services were employed to detect matches including MatInspector (Quandt *et al.*, 1995), SignalScan (Prestridge, 1991) and PLACE (Higo *et al.*, 1998). These search engines use different but overlapping datasets of published transcription factor binding sites including TFD (Ghosh, 1991) and TRANSFAC (Heinemeyer *et al.*, 1998). They also differ in the search algorithms which allow in some cases for near matches to be reported (Higo *et al.*, 1998; Quandt *et al.*, 1995).

Several exact matches to consensus sequences for *cis*-acting elements and plant transcription factor binding sites were identified. To simplify the analysis of the sequence the matches were subjected to a selection procedure based on the likelihood of their occurrence in the sequence. Specifically, the probabilities of the random occurrence of the observed number of hits of each match in any sequence with the length and composition characteristics of that analyzed were calculated. Calculations were performed with the help of a program written for the effect. Those matches for which the number of hits had an associated probability of less than 0.4 are shown in Figure 8.6. Included were a TATA box-like sequence (TATAT-ATAAT) 28 bp upstream from mP18. Other matches to the TATA-box consensus (TATAWAW) were found at positions -375 (TATAAAT) and -770 (TATATAT) relative to the 5' end of mP18. Several matches to other transcription factor binding sites were also found throughout the region but were mostly clustered closely upstream from mP18 (Fig. 8.6). Those sites found within 350 bp upstream from mP18 are listed in Table 8.1. They included an ASF-1 motif (TGACG) at -77 relative to mP18, and a L box (TCCTACCAACC) at -135 overlapping an H-box (CCTACC) at -136, all in the complementary strand. The L box was preceded by an I-box (GATAAG) at -152. A second H-box, in the complementary strand, was located at -171. Two GT-1 boxes separated by 21 bp were present at -279 and -300, and a match to a 7S globin box (Lessard *et al.*, 1991) was found nearby at -338.

For each element listed in Table 8.1, the probability of the observed number of hits occurring in the 350 bp sequence upstream from mP18 was calculated taking in consideration the composition of the sequence. Clearly, some matches have a low probability of being present in a sequence with the composition found in the region, including the 7S globin box ($P=1.015 \times 10^{-2}$), the L box ($P=2.528 \times 10^{-4}$), and the H-boxes (3.471×10^{-4}).

Name/Position	Sequence/Strand	Consensus	Reference	Probability
7S globin box -338	ATATTTAAT (+)	ATATTTAWW	(Lessard <i>et al.</i> , 1991)	1.015×10^{-2}
I-box core -300	GATAA (+)	GATAA	(Terzaghi & Cashmore, 1995)	0.364
H-box -171 -136	CCTACC (-) (-)	CCTACC	(Terzaghi & Cashmore, 1995)	3.471×10^{-4}
I-box -158	GATAAG (+)	GATAAG	(Donald & Cashmore, 1990)	0.151
L-box -135	TCCTACCAACC (-)	YCYYACCWACC	(Logemann <i>et al.</i> , 1995)	2.528×10^{-4}
E-box -91 -1	CAAATG (+) CATTTG (+)	CANNTG	(Stalberg <i>et al.</i> , 1996)	0.2714
ASF-1 motif -77	TGACG (-)	TGACG	(Terzaghi & Cashmore, 1995)	0.321
Hex motif H3H4 -77	ACGTCA (+)	ACGTCA	(Mikami <i>et al.</i> , 1989)	0.183
TATA box 4 -28	TATATAA (+)	TATATAA	-	0.217
GT-1 -300 -279	GATAAT (+) GAAAAA (+)	GRWAAW	(Terzaghi & Cashmore, 1995)	0.248

Table 8.1-Matches to consensus *cis*-regulatory elements of transcription in plants.

These matches were found in the 0.35 kb segment upstream from mP18. The distance to the first base of mP18 (defined here as +1) is indicated, as is the strand where the match occurred (+ or -). Also shown are the adopted consensus sequences as defined in the various databases. The original references for the motifs are indicated. For each consensus motif the probability of occurrence of the observed number of matches within the 350 bp sequence was calculated and is shown in the last column.

Further upstream several matches to other types of regulatory elements were observed. Included were three RRCCCR elements at -602, -646, -753 ($P_3 = 2.439 \times 10^{-5}$) and a second match to the 7S globin box at -494 ($P_2 = 6.985 \times 10^{-2}$).

8.3 Sizing the fusion transcripts by Northern analysis

The analysis of the genomic region upstream of mP18 revealed the possible existence of a single additional exon located between 137 and 39 bp upstream from mP18 (Section 8.2.4). The distance separating this hypothetical element from mP18, 39 bp, was below the known limit for intron length in plants (Brown *et al.*, 1996). The 5' RACE of the fusion transcripts also failed to detect transcripts starting further upstream from mP18 (Section 6.2). In addition the 5' end of mP18 is closely preceded by a TATA-box, whereas the two other TATA-boxes found in the region were considerably further upstream at 0.4 kb and 0.8 kb. In AtEN101 the 5' RACE products and the structure of the T-DNA suggested that fusion transcripts from the left junction should be about 2.5-2.8 kb depending on the precise site of polyadenylation and the size of the polyA⁺ tail. This was considerably shorter than the fusion transcripts detected by Topping *et al.* (1994). However, the longer 5' RACE products detected (Section 6.2.4) were less abundant and apparently not linked to a *gus* gene (Section 6.3). These observations suggested that the fusion transcripts might be shorter than had previously been estimated (Topping *et al.*, 1994). It was therefore of interest to reassess the size of these transcripts in AtEN101.

To determine the size of the fusion transcripts samples of previously prepared total RNA of C24 and AtEN101 (Section 2.6.2) were run in a denaturing formaldehyde-agarose gel (Fig. 8.7A). The gel was blotted onto a nylon membrane and this Northern blot was hybridized with a probe prepared from the *gus* gene. The probe, labelled with [α -³²P]dCTP, was obtained by the usual procedure (Section 2.8.1) from the *Bam*H I-*Eco*R I insert of pPRGUS-6 encompassing the entire length of the *gus* gene and associated *nos* terminator. The hybridization was in formamide solution at 48°C, and at the end of the incubation period the excess probe was removed by washing the Northern blot, as described in Section 2.8.4. For detection of transcripts hybridizing to the probe a Phosphoimager screen was exposed to the Northern blot for 3 weeks. At the end of this period the screen was scanned in a Phosphoimager detector (Section 2.8.5). As can be seen in Figure 8.7B, a band of transcripts of about 2.7-2.8 kb hybridized to the *gus* probe. Given the known structures of the left and right junctions in AtEN101 (Section 6.3), these transcripts were expected to originate close to the T-DNA insertion site (up to about 0.2 kb). This prediction was in good agreement with the sizes of the products of 5' RACE of fusion transcripts (Section 6.2).

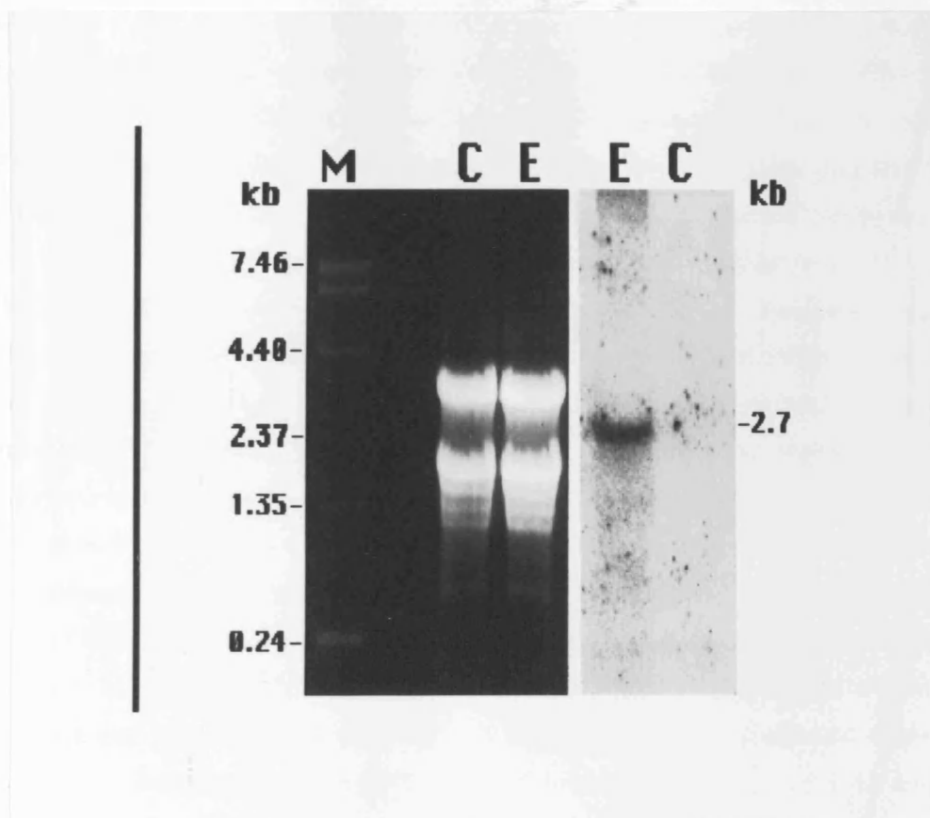


Figure 8.7 Sizing the fusion transcripts in AtEN101 by Northern blot analysis.

- A)** Aliquots of total RNA of C24 (20 μ g; Lane C) and AtEN101 (35 μ g; Lane E) were resolved in a 1% formaldehyde-agarose denaturing gel. An RNA ladder marker (Gibco/BRL) was run in parallel (Lane M). A Northern blot of the gel shown was subsequently prepared (Section 2.7.2).
- B)** Detection of fusion transcripts by hybridization. The Northern blot of the gel shown above was hybridized with a probe prepared from the *gus* gene and labelled with [α - 32 P]dCTP. The hybridization was performed in 50% formamide at 48°C, as described in Section 2.8.4. For detection of hybridizing transcripts the blot was placed in contact with a Phosphorimager screen for 3 weeks. Scanning of the screen in a Phosphorimager detector resulted in the image shown. Lane E: AtEN101 sample; Lane C: C24 sample.

8.4 The promoter region of *PKT2*

8.4.1 Searching for *PKT2* transcripts in AtEN101

The existence of *PKT2* transcripts in AtEN101, where the T-DNA is interposed in the genome 0.3 kb upstream from mP19, would be compatible with the possible presence of the promoter of the gene in this segment. To test the presence of *PKT2* transcripts in AtEN101 a Northern blot of RNA obtained from different organs was hybridized with the probe prepared from the insert of pCEN1-8. After a long period (3 weeks) of exposure, an autoradiogram was obtained where very faint hybridization was observed in the region corresponding to the 16S chloroplast rRNA (data not shown). Because the results were unclear, the more sensitive method combining reverse transcription and PCR, RT-PCR, was used for detection of *PKT2* transcripts. Reverse transcription was performed on total RNA isolated from mature leaves with primers prTHIO-5R and LB-267. Primer prTHIO-5R represents a segment of transcript overlapping two exons, eliminating possible contamination by genome amplification. Two sets of PCR amplification reaction were performed, including a positive control testing for the presence of fusion transcripts at the left junction with primers LB-267 and prTHIO-3F. The other used prTHIO-5R together with prTHIO-4F. A first round of amplification for 30 cycles was insufficient to detect any *PKT2* transcripts. A second round for 20 cycles was then performed, but with similar results (Fig. 8.8). In contrast, a band of fragments of about 0.19 kb were detected, as expected, in the positive control reaction. Repetitions of the test also failed to detect *PKT2* transcripts.

Although *PKT2* transcripts were not detected by RT-PCR, their absence is not certain. The presence of the T-DNA in the putative promoter region might have caused a displacement of the transcription initiation site(s) to a location downstream from the prTHIO-4F segment. Another possibility is that the regulation of transcription was altered so that it may only occur in other organs or under different physiological conditions.

8.4.2 Matches to *cis*-regulatory elements in the promoter region of *PKT2*

The failure to find matches to splice site consensus sequences in the 623 bp-long segment comprised between the 5' ends of mP18 and mP19 (Section 5.3.5) was in agreement with the model of two overlapping genes. Analysis of the segment with NetPlantGene V2.1 (Hebsgaard *et al.*, 1996), and Grail 2 (Xue *et al.*, 1994), confirmed the earlier results. This segment was therefore expected to include the promoter of *PKT2*.

The 623 bp-long segment was inspected for matches to known *cis*-acting regulatory elements of plants. The search was performed with MatInspector (Quandt *et al.*, 1995), SignalScan (Prestridge, 1991) and PLACE (Higo *et al.*, 1998), as for the promoter region of *PKT1* (Section 8.2.3). Several matches to cognate binding sites for transcription factors were found in the region (Fig. 8.9). The CCAAT box previously found upstream from the

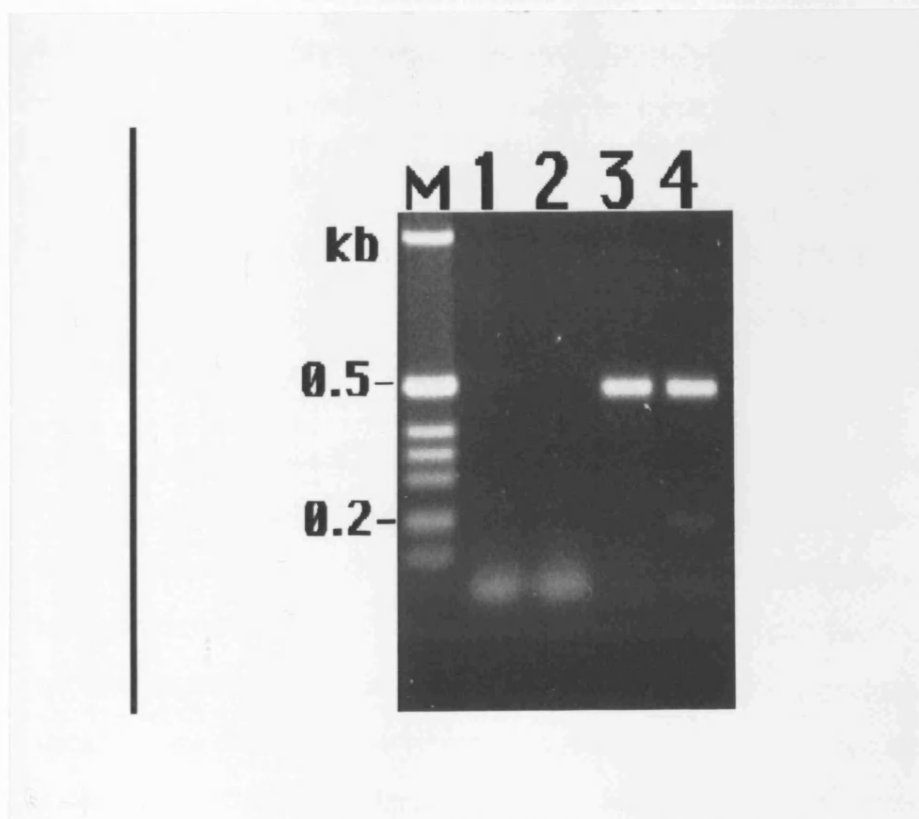


Figure 8.8 Testing the presence of *PKT2* transcripts in AtEN101.

The possible presence of *PKT2* transcripts in leaves of AtEN101 homozygous plants was investigated by RT-PCR. Reverse transcription was performed with primer prTHIO-5R and, for a positive, control with LB-267. Amplifications were with the primer pair prTHIO-5R/prTHIO-4F and, for the positive control, LB-267/prTHIO-3F. A first round of amplification was performed for 30 x (94°C 25 secs, 54°C 20 secs, 72°C 35 secs). A second amplification was initiated with 10% of the first reaction and lasting for 20 x (94°C 25 secs, 54°C 20 secs, 72°C 35 secs). Lane 1: prTHIO-5R/prTHIO-4F; Lane 2: prTHIO-5R/prTHIO-4F (negative control of RT); Lane 3: LB-267/prTHIO-3F (negative control); Lane 4: LB-267/prTHIO-3F.


```

-663 TGTATATATAATCTTGAACAATCGGTTATCTCAACATTTGCATATTGTCTACTTTTAGAT
      E5.2
-603 TTTGATATCTATTGGAACTCTTACTAGCTCATTTCATCCGTCGGGTAACCTTTCTTGACCTC
      ---1---      ---4---
      ---2---
      ---3---
-543 TCTCCATCTTATCTTTTACGAGTCTTAAAGACTTTACTTATCTAAGAAAATCCGTTTTAG
      --6---      --6---      --6---
-483 TCTCTCAAAATTGATTTATTAGATGGTATATTCTAGTCACAACAACCTATTTTTTGGACGA
      --7---
-423 GAATTTTACAAAATAATTGAAGGATACTTAACTCTTTCTTACTAAACAAAACAACCTTAAA
      E1.5      --8---
-363 TCTATATCCGATATGAATTCTTGTGGGTAAAGAATACAAATCATGTATATTATCTGGGATA
      --8---      --6---      --6---      --8---
      ---
-303 TCCAAATTATTCACTAAAATTTTATAAATATCTCAAATTCTTCTAAATTTGTAAATTAC
      8--      --6---
-243 TCAAATTATCAAAATCACTAATACCCAAAATGAGAAGAAACATTTATTGCTTAAAAGTTA
      --6---
-183 GATTCTTCCCACGTGGCCCGGTTTAGTTATAGTTGCCGACGATTAATTCCCGGTCACATT
      ---9---      --2---
      --3---
-123 TTTAACGGCGTGATTTGTATGTGGCGTAAAACATGTCAAATGCGATATTTTCGTAACGA
      --6---      -2-
-63 CGTTAACTTATACAAATACGACAAGACAAGTTTTTGGCAACGAGGCCAATCAGACCAGAC
      --
-3 TTGTGAGAACTGAGAACGATCGATCAAAAATCTCTCCCACGATAAAAAATGAGGACCCTGA
64 AGCTCATTGTACTTAACGTCTTCCCTTTCCAGAAAAATCAACCAAACCCATTTTCAGTTT
124 CCCCCGAAAGTTTTTGTCTTTCAATCACCATTAGAATTTCAGCTGCAACACAAGTGTTC
      E1.4
184 AACAGAGAGAAATCATGGAGAGAGCTATGGAAAGACAAAAGATATTGCTTCGTCATCTCAA
244 TCCAGTTTCTTCTTCTAATTCTTCTCTTAAACATGAACCTTCTTCTGTCTGTAAGTTT

```

Figure 8.9 Matches to transcription factors binding sites in putative promoter region of *PKT2*. Shown is a 960 bp genomic segment comprising the first exons of mP18 (colored red) and mP19 (colored green). The match to the TATA-box consensus is highlighted by double underlining. Matches to consensus *cis*-regulatory elements listed in Table 8.2 for the *PKT1* intron segment upstream from mP19 are shown in blue. They are identified by the numbered undelining: 1) Hexameath4; 2) Amy3 box; 3) LTRE box; 4) ELRE core; 5) I-box; 6) GT-1 site; 7) SF1-box; 8) Myb-st1. Where two elements overlap, the smaller is underlined. Italics indicate matches occurring in the complementary strand. The two *EcoR* I sites E1.4 and E1.5 are labeled and underlined.

pCEN1-9 cDNA (Section 5.4.1) was located at only 13 bp from mP19. Among the elements found, those with a lower probability of occurring in a sequence with a base composition of the segment analyzed are listed in Table 8.2. A match to the TATA-box consensus (TATAWAW), TATAAAT, was present 275 bp upstream from the mP19 5' end. Other sequences of interest included CCCACGTGGCC at -165 which encompasses several matches to ABRE elements (see Chapter 11), among which YACGTGGC is known to be a strong ABRE element (Busk & Pages, 1998). In addition, this sequence contained a match to the G-box core consensus (CACGTG) (Menkens *et al.*, 1995), an EMBP-1box (Guiltingan *et al.*, 1990), and an E-box (Stalberg *et al.*, 1996). Six matches to the GT-1 box were dispersed throughout the sequence but four of these were located within a 105 bp-long segment at position -233, three of which upstream from the TATA box. Two I-boxes separated from each other by 23 bp were located at -502. Other sequence matches are listed in the Table 8.2. The likelihood of occurrence of most of the sequence elements listed was low, particularly those of the G-box ($P_1 = 5.228 \times 10^{-4}$), Amy3 ($P_3 = 7.414 \times 10^{-4}$), GT-1 sites ($P_6 = 1.482 \times 10^{-3}$) and Myb-st1 ($P_4 = 9.455 \times 10^{-3}$) matches.

8.5 Functional analysis of the putative promoter regions

8.5.1 Rationale

The structural analysis indicated that the promoters of *PKT1*, *PKT2* and for the fusions transcripts should be located closely upstream from the 5' termini of mP18 and mP19. To delimit more precisely the promoter of both native genes a functional analysis strategy was adopted. Various segments upstream of the transcribed sequences mP18 and mP19 were fused to a promoterless *gus* reporter gene previously cloned into the vector pBluescript. The capacities of these constructs to drive transcription of the reporter gene were then evaluated by transient expression *in vivo* after delivery to target plant material by particle bombardment. The potential of the region containing the putative promoter of *PKT2* to drive the expression of the *gus* gene located at the T-DNA right junction in AtEN101 was similarly tested by transient expression. For testing, a construct was assembled comprising the segment of interest fused to a downstream *gus* gene in the plasmid vector pBluescript.

8.5.2 Construction of the test plasmids.

8.5.2.1 Putative promoter region of *PKT1*

For testing the promoter of *PKT1* four segments of the genome were selected. These segments were derived from pTNB8-x clones by digestion with *EcoR* V for which there is a single site

Name/Position	Sequence/Strand	Consensus	Reference	Probability
Hexamer ath4 -562	CCGTCG (+)	CCGTCG	(Chaubet <i>et al.</i> , 1996)	2.225×10^{-2}
Amy3 -563 -143 -62	CGTCG (-) CGACG (+) CGACG (+)	CGACG	(Hwang <i>et al.</i> , 1998)	7.414×10^{-4}
LTRE -561 -144	CCCGA (-) GTCGG (+)	GTCGG	(Jiang <i>et al.</i> , 1996)	1.195×10^{-2}
ELRE core -546	TTGACC (+)	TTGACC	(Rushton <i>et al.</i> , 1996)	9.926×10^{-2}
I-box -531 -502	CTTATC (-) (-)	GATAAG	(Donald & Cashmore, 1990)	3.089×10^{-2}
GT-1 -494 -333 -310 -270 -234 -71	GAAAAT (+) GGTAAA (+) ATTATC (-) GATAAT (-) ATTATC (-) ATTTTC (-)	GRWAAW	(Terzaghi & Cashmore, 1995)	1.482×10^{-3}
S1F-box -456	ATGGTA (+)	ATGGTA	(Zhou <i>et al.</i> , 1992)	0.135
Myb-st1 -399 -355 -304 -301	GGATA (+) TATCC (-) GGATA (+) TATCC (-)	GGATA		9.455×10^{-3}
TATA-box 2 -275	TATAAAT (+)	TATAAAT	(Shirsat <i>et al.</i> , 1989)	0.238
G-box -167	CCACGTGGC	MCACGTGGC	(Busk & Pages, 1998)	5.228×10^{-4}

Table 8.2 Matches to *cis*-acting regulatory elements of transcription of plants found upstream from mP19.

These matches were found in the 623 bp segment upstream from mP19. The distance to the first base of mP19 (defined here as +1) is indicated, as is the strand where the match occurred (+ or -). Also shown are the adopted consensus sequences as defined in the various databases. The original references for the motifs are indicated. For each consensus motif the probability of occurrence of the observed number of matches within the sequence was calculated and is shown in the last column.

in the TnΔXR sequence at 168 bp from the δ terminus. The clones were chosen for the location of their transposon at different distances upstream from mP18 and, consequently, from the only two *EcoR* V sites in the B1.4B1.5 segment (cf. Section 8.2.3; Figure 8.5). Additionally, to uniformize and minimize the contribution of TnΔXR sequences to the

constructs, only clones where the proximal side of the transposon relative to mP18 was the δ terminus were selected. As a result, the segments tested had as common 3' terminus (in the direction of transcription) the midpoint of the E5.1 site, located 26 bp from the 5' end of mP18. At their 5' ends they also shared the 168 bp derived from the segment between the midpoint of the *EcoR* V site in Tn Δ XR and its δ terminus. The clones chosen were pTNB8-28, -72, -66, and -31, which contained a total of 386 bp, 685 bp, 894 bp, and 1266 bp of *Arabidopsis* genomic sequences, respectively.

Following digestion of the selected pTNB8-x clones and resolution of the products by agarose gel electrophoresis (Fig. 8.10A), the relevant fragments were isolated by electroelution and purified by the standard method (Section 2.4.6.1). Samples of re-dissolved fragments were run in a gel to assess their quality (Fig. 8.10B). Each of the purified fragments was then linked to the promoterless reporter gene in plasmid pPRGUS-6. This plasmid carries a copy of the *gus* gene and an associated *nos* terminator as a *Bam*H I-*Eco*R I insert in vector pBluescript SK(+). For assemblage of the constructs pPRGUS-6 was linearized with *Sma* I, which digests at a single site in its polylinker, 16 bp upstream from the start codon of *gus*. Prior to ligation to the pTNB8-x *Eco*R V fragments the linearized plasmid was dephosphorylated with calf intestinal phosphatase and purified. *E. coli* was transformed with the products of the ligations and several transformants isolated for each. These clones and their plasmids were named pPRP18.x-y according to the promoter region they should correspond to (mP18) and the pTNB8-x plasmid from which the specific fragment derived.

To select suitable transformants, minipreparations of plasmid DNAs were obtained for several of them. The plasmids were subjected to restriction enzyme analysis to establish the presence of the genomic segment in them and its orientation relative to the reporter gene (*e.g.* Fig. 8.11). For diagnostic purposes *Sac* I, and *Hinc* II plus *Sac* I were the enzymes chosen given the restriction maps predicted for the recombinant plasmids. The clones containing the plasmids selected for subsequent work were pPRP18.28-106, pPRP18.31-3, pPRP18.66-3, and pPRP18.72-105. The structural features of these plasmids are exemplified in the maps shown in Figure 8.12 along with a restriction enzyme map of pPRP18.72-105. Considering that the plasmids were destined to be used in transient expression assays, it was important to ensure that the *Arabidopsis* genomic sequences were intact within them, particularly nearer the mP18 5' end. For this reason automated sequences in a single or both strands were obtained for each of the plasmids with the universal reverse primer and the Nested 2 primer (Lindsey *et al.*, 1993). The sequence for the latter primer is located 8 bp downstream from the *Sma* I site in pPRGUS-6, in the direction of the 5' end of mP18 from which it is separated by 37 bp. No differences to the wildtype genomic sequence were observed in any of the four clones.

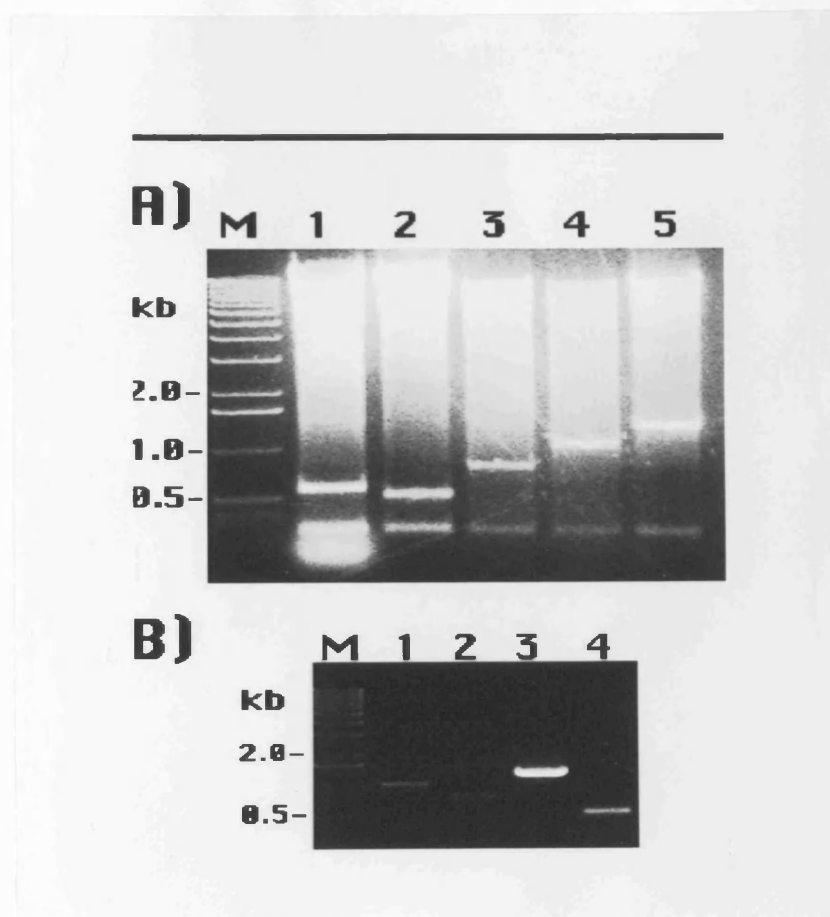


Figure 8.10 Isolation of segments from the promoter region of *PKT1* in pTNB8-x plasmids.

- A)** *EcoR* V digests of selected plasmids were separated in agarose gels (an example is shown). The bands corresponding to the presumed promoter region were dissected out of the gels and recovered by electroelution. Loading order was: Lane M: 1 kb DNA marker ladder; Lane 1: pTNB8-3 (not used for constructs); Lane 2: pTNB8-28; Lane 3: pTNB8-31; Lane 4: pTNB8-72; Lane 5: pTNB8-66. Because of the heavy chromosomal contamination visible in this gel, the DNA was purified from digests of other plasmid preparations resolved in similar gels (data not shown).
- B)** Samples of the purified *EcoR* V fragments were resolved in an agarose gel to verify their quality and estimate their concentration for subsequent ligations. Lane M: 1kb DNA ladder marker. Lane 1: promoter fragment from pTNB8-66; Lane 2: from pTNB8-72; Lane 3: from pTNB8-31; Lane 4: from pTNB8-28.

Figure 8.11 Diagnostic digests of pPRP18.x-y constructs.

Top: Aliquots of *Sac* I digests of the various pPRP18.x-y and pPRGUS-6 plasmids were resolved in agarose gels. Lanes M: 1 kb DNA ladder marker (Promega); Lane 1: pPRP18.28-106; Lane 2: pPRP18.31-3; Lanes 3 and 4: pPRP18.66-3; Lane 5: pPRGUS-6. The common 3.2 kb fragments of the pPRP18x-y plasmids containing the vector and the *nos* terminator are visible as are the 0.1 kb partial segment of the Tn Δ XR transposon (please refer to Figure 8.12 for restriction and features map).

Bottom: Aliquots of the pPRP18-x-y and pPRGUS-6 plasmids digested *Sac* I plus *Hinc* II were submitted to electrophoretic separation in agarose gels. Lanes M: 1 kb DNA ladder marker (Promega); Lane 1: pPRP18.28-106; Lane 2: pPRP18.31-3; Lane 3: pPRP18.72-15; Lane 4: pPRGUS-6; Lane 5: pPRP18.66-3.

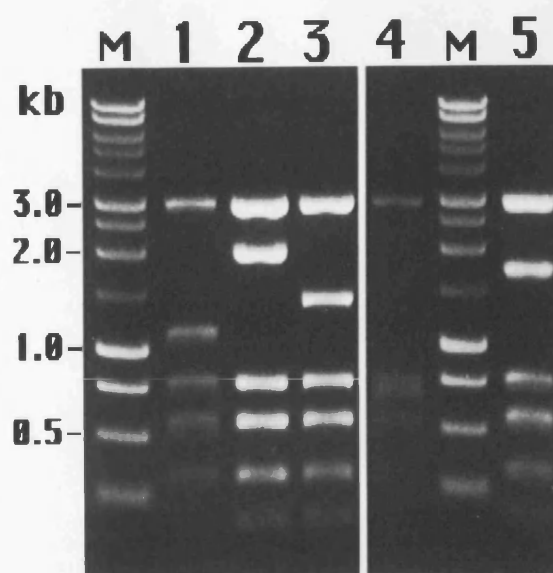
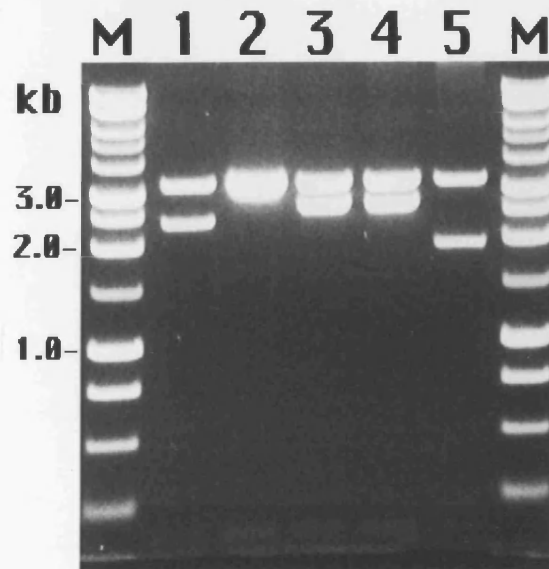
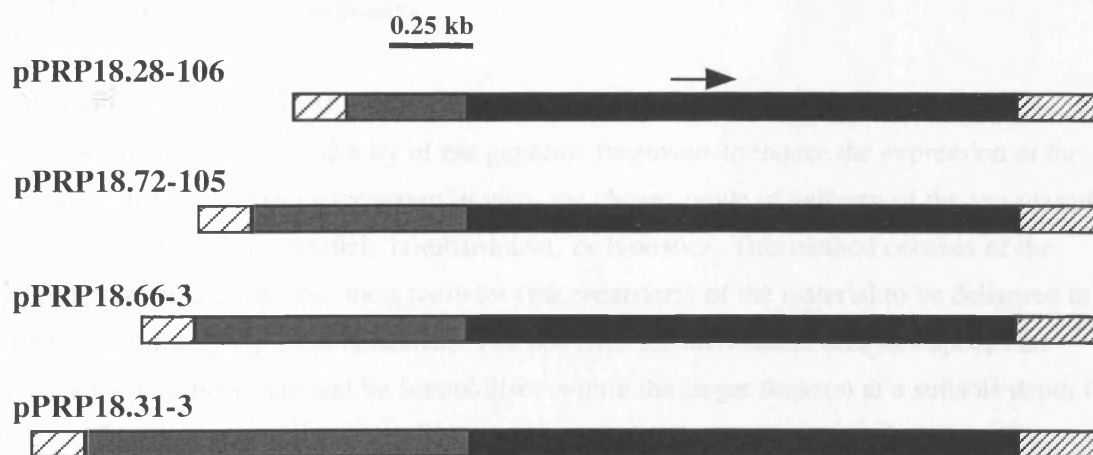
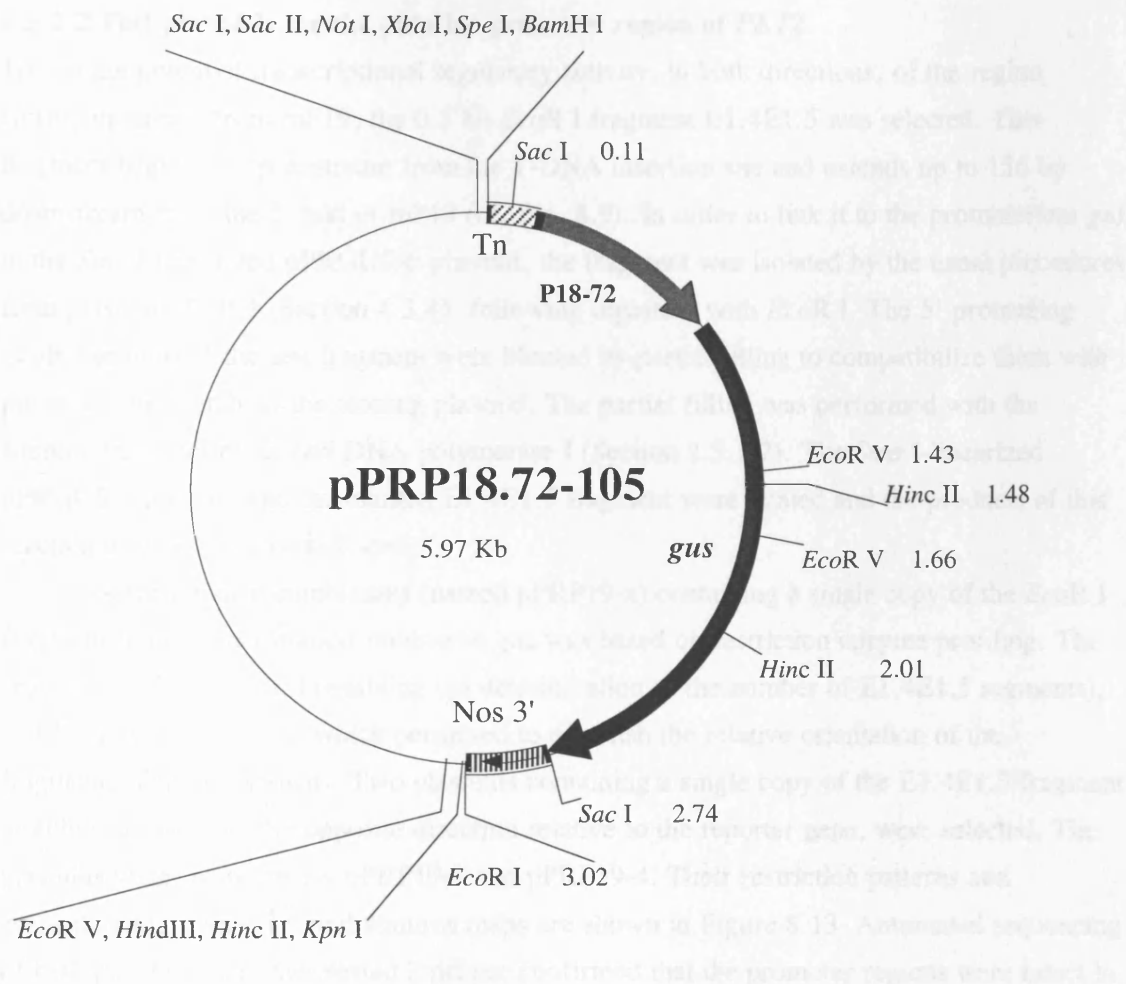


Figure 8.12 Features and restriction maps of pPRP18.x-y plasmids.

Top: A combined features and restriction map for pPRP18.72-105 is shown.

Bottom: Features maps for the insert regions of all four selected pPRP18.x-y plasmids. The direction of the reporter gene *gus* is indicated by the arrow.



Legend

- | | |
|--------------------------------|--------------------------|
| TnΔXR | <i>gus</i> reporter gene |
| <i>A. thaliana</i> genomic DNA | <i>nos</i> terminator |

8.5.2.2 Test plasmids for the putative promoter region of *PKT2*

To test the potential transcriptional regulatory activity, in both directions, of the region closely upstream from mP19, the 0.5 kb *EcoR* I fragment E1.4E1.5 was selected. This fragment begins 12 bp upstream from the T-DNA insertion site and extends up to 156 bp downstream from the 5' end of mP19 (cf. Fig. 8.9). In order to link it to the promoterless *gus* in the *Sma* I-linearized pPRGUS-6 plasmid, the fragment was isolated by the usual procedures from plasmid pENE2 (Section 4.3.4), following digestion with *EcoR* I. The 5' protruding *EcoR* I termini of the test fragment were blunted by partial filling to compatibilize them with the blunt *Sma* I ends of the cloning plasmid. The partial filling was performed with the Klenow fragment of *E. coli* DNA polymerase I (Section 2.5.1.2). The *Sma* I-linearized pPRGUS-6 plasmid and the blunted E1.4E1.5 fragment were ligated and the products of this reaction used to transform *E. coli*.

Selection of recombinants (named pPRP19-x) containing a single copy of the *EcoR* I fragment in either orientation relative to *gus* was based on restriction enzyme profiling. The enzymes used were *Sac* I (enabling the determination of the number of E1.4E1.5 segments), and *EcoR* V and *Hinc* II (which permitted to establish the relative orientation of the fragments; data not shown). Two plasmids containing a single copy of the E1.4E1.5 fragment in either the same or the opposite direction relative to the reporter gene, were selected. The plasmids were, respectively pPRP19-3 and pPRP19-4. Their restriction patterns and corresponding restriction and features maps are shown in Figure 8.13. Automated sequencing of both plasmids with the Nested 2 primer confirmed that the promoter regions were intact in them.

8.5.3 Transient expression assays

8.5.3.1 Fundamentals of the assay method

To evaluate the potential capacity of the genomic fragments to induce the expression of the reporter gene by transient expression *in vivo*, the chosen mode of delivery of the test plasmids to the living tissues was particle bombardment, or biolistics. This method consists of the deposition over microscopic inert particles (microcarriers) of the material to be delivered to the interior of the biological specimen. The particles are then accelerated to a speed that enables them to penetrate and be immobilized within the target tissue(s) at a suitable depth for uptake of the test material.

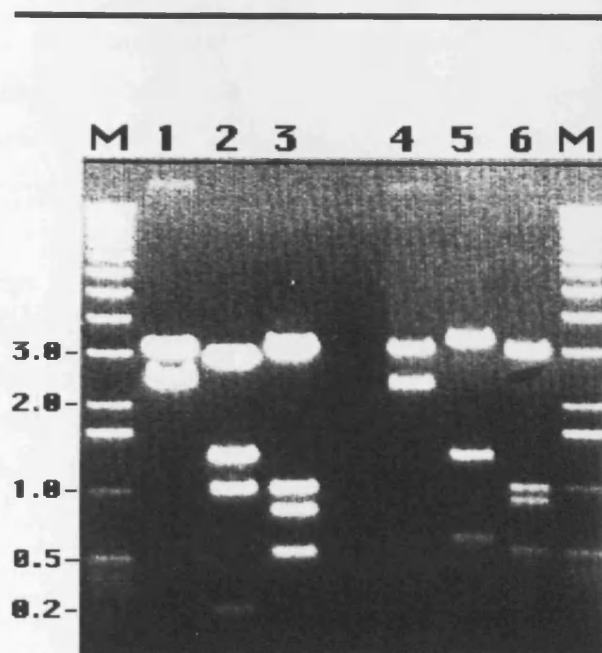
Several variant designs exist for biolistic systems. In the present case a BIOLISTIC[®] PDS-1000/He (Bio-Rad) was used, as described in detail in Section 2.9. In this design the microcarriers are loaded onto a macrocarrier that is mounted on the BIOLISTIC[®] PDS-1000/He. The target is also placed in the chamber of the apparatus and a gaseous partial

Figure 8.13 Organization of the pPRP19-3 and pPRP19-4 constructs.

A) Restriction analysis of the plasmids. *EcoR* V, *Hinc* II and *Sac* I digests were electrophoresed in an agarose gel. Lanes M: 1 kb DNA ladder marker. Lanes 1-3: pPRP19-3 digested with *Sac* I, *EcoR* V, and *Hinc* II. Lanes 4-6: Similar digests of pPRP19-4. The analysis demonstrated that the plasmids had the structures shown below.

B) Restriction and features maps of pPRP19-3 and pPRP19-4.

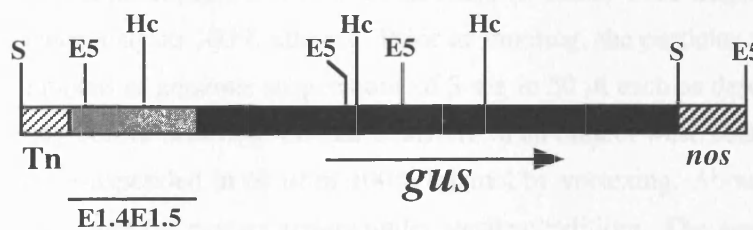
A)



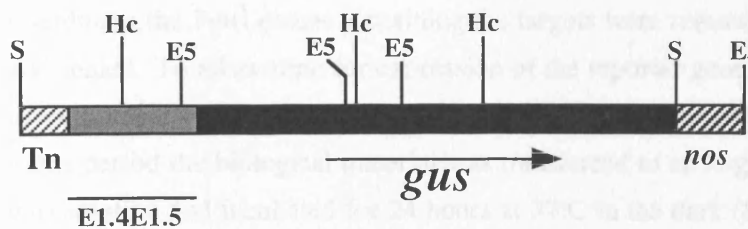
B)

pPRP19-3

0.5 kb



pPRP19-4



vacuum made. A sudden burst of helium gas fired into the chamber causes the acceleration of the macrocarrier and propels it from the rest position towards the target. The movement of the macrocarrier is halted by a stopping screen which allows the microcarriers to proceed by inertia and penetrate the biological target. The internal pressure of the chamber is allowed to equilibrate with atmospheric pressure, the material removed and subjected to the pertinent analysis. In the present case, the tissues were tested histochemically for GUS activity (Section 2.10.2).

8.5.3.2 Transient assays

The effectiveness of the transient assays is dependent on several factors including the penetrative power of the microcarriers, their dispersal, and the nature of the target material. In a first set of experiments these factors were optimized in a qualitative manner. The target biological materials utilized were fresh young leaves of tobacco SR1 plants grown in sterile MS30 media (Section 2.1.4.4). A plasmid containing the reporter gene under the control of the 35S CaMV promoter, pMKC1 (a generous gift of Mr. M.-K. Cheung), was employed as a positive control of *gus* expression. This permitted the qualitative determination of an effective distance between the macrocarrier stopper and the target leaves, as well as a suitable burst pressure for the helium shot (1350 or 1550 psi). Combined, these parameters resulted in good dispersal of the microcarriers and signal upon detection of GUS activity (data not shown).

For the assays involving the constructs of the *PKT1* and *PKT2* promoter regions, the original reporter gene-carrying vector, pPRGUS-6, was used as the negative control. Two large-scale preparations were performed for all constructs and pPRGUS-6 by the method described in Section 2.4.1.2 in order to obtain high quality supercoiled plasmids, carrying little bacterial chromosome contamination. Gold microcarriers (BioRad) were washed and sterilized by vortexing vigorously in 100% ethanol. Prior to shooting, the particles were pelleted, washed and aliquoted as aqueous suspensions of 3 mg in 50 μ l each as described (Section 2.9). Immediately before shooting the microcarriers in an aliquot were coated with 5 μ g of DNA, pelleted and resuspended in 60 μ l of 100% ethanol by vortexing. About 10 μ l of the suspension were loaded onto the macrocarriers under sterile conditions. The macrocarriers and rupture disks were mounted and the target material placed in its holder prior to shooting. For each shot, one or more freshly excised leaves of tobacco SR1 plants were used as targets after their placing with their adaxial surface down on 1/2MS agar medium in a Petri dish.

After each bombardment the Petri dishes containing the targets were removed from the chamber, covered and sealed. To allow time for expression of the reporter gene, the bombarded specimens were incubated in the Petri dish at 25 °C on a day/night cycle for 18 to 24 hours. At the end of this period the biological material was transferred to an X-gluc (Biosynth G, Switzerland) solution and incubated for 24 hours at 37°C in the dark (Section 2.10.2). The specimens were then cleared by passing them through an ethanol series up to 95

%. This permitted the unhindered visualization of spots of indigo color indicative of GUS activity.

Testing of each construct was performed on three or four occasions spaced by several days or weeks, using different batches of plants and either of the two plasmid preparations. Each test of a construct included between two and four bombardments. Every bombardment used a different target usually with a burst pressure 1550 psi. Negative control assays were performed in every testing session with pPRGUS-6. The results of the tests are summarized in Table 8.3. The observations associated with each construct were all consistent. GUS activity was detected in all assays of constructs of the putative promoter regions. No enzymatic activity was found in the assays of pPRGUS-6, or in two other assays conducted with uncoated (sterilized) microcarriers. Figure 8.14 (A-F) shows examples of the results obtained with leaves bombarded with each of the constructs.

With the intent of testing the ability of the constructs to drive *gus* expression in developing plant embryos, some assays were also performed by bombarding developing seeds of *Phaseolus vulgaris*. To expose the endosperm and developing embryos prior to bombardment, the immature seeds were longitudinally dissected in two halves under sterile conditions. The seeds were then placed with their cut surfaces up on plates of 1/2MS agar medium and bombarded. Because of practical difficulties in manipulating the material and in maintaining sterile conditions the *Phaseolus* seeds were not utilized in a systematic manner. Of the set of test constructs only plasmid pPRP18.72-105 and the negative control pPRGUS-6 were used in these transient assays (Table 8.3). GUS activity was found both in the endosperm and embryos as well as maternal tissues of seeds bombarded with pPRP18.72-105 (Fig. 8.14G), but not those targeted with pPRGUS-6.

8.6 Summary

In the previous Sections the regions presumed to contain the promoters of *PKT1* and *PKT2* were structurally and functionally analyzed. The sequence of about 1.2 kb upstream from mP18 was determined (Section 8.2). This involved the generation of a family of insertion derivatives of pENB8 with transposon TnΔXR, their selection and sequencing. Analysis of the sequence failed to detect any additional exons of the *PKT* genes. Several matches to *cis*-acting regulatory elements of transcription were found in the vicinity of mP18, including a TATA-box 28 bp upstream from it. The results of the structural analysis together with those of the 5' RACE of fusion transcripts (Section 6.2) led to a re-evaluation of the size of the latter by Northern hybridization (Section 8.3). Fusion transcripts of 2.7-2.8 kb were detected, in agreement with the possible co-localization of their 5' ends with those of the *PKT1* transcripts.

pPRP18.28-105		pPRP18.72-106		pPRP18.366		pPRP18.31-3		pPRP19-1		pPRP19-4		pPRGUS-6	
Assay	GUS	Assay	GUS	Assay	GUS	Assay	GUS	Assay	GUS	Assay	GUS	Assay	GUS
1-1 (A)	+	1-1(A)	+					1-1 (A)	+			1-1 (A)	-
1-2 (A)	+	1-2 (A)	+					1-2 (A)	+			1-2 (A)	-
		1-3 (A)	+									1-3 (B)	-
												1-4 (B)	-
2-1 (A)	+	2-1 (A)	+									2-1 (A)	-
2-2 (A)	+	2-2 (A)	+									2-2 (A)	-
2-3 (A)	+	2-3 (A)	+										
						3-1 (A)	+	3-1 (A)	+			3-1 (A)	-
						3-2 (A)	+	3-2 (A)	+			3-2 (A)	-
						3-3 (A)	+	3-3 (A)	+				
4-1 (B)	+			4-1 (A)	+	4-1 (A)	+					4-1 (A)	-
4-2 (B)	+			4-2 (A)	+	4-2 (A)	+					4-2 (A)	-
				4-3 (A)	+	4-3 (A)	+						
						5-1 (B)	+	5-1 (B)	+	5-1 (A)	+	5-1 (B)	-
						5-2 (B)	+	5-2 (B)	+	5-2 (A)	+	5-2 (B)	-
						5-3 (B)	+	5-3 (B)	+	5-3 (A)	+		

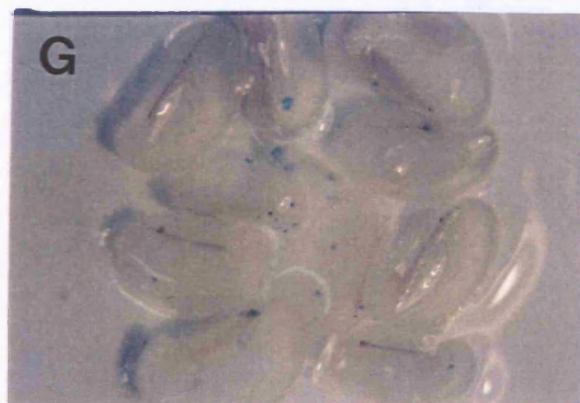
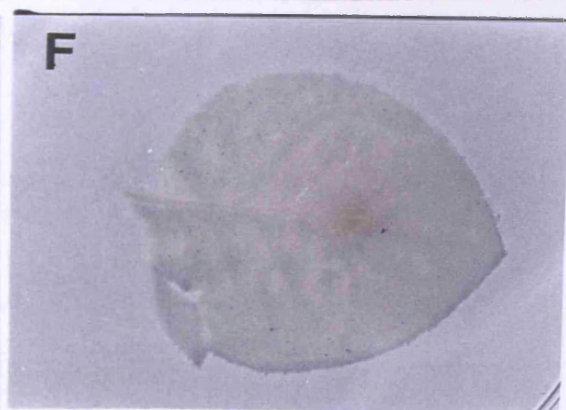
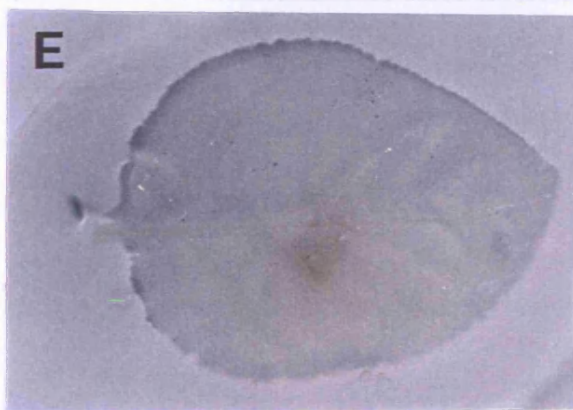
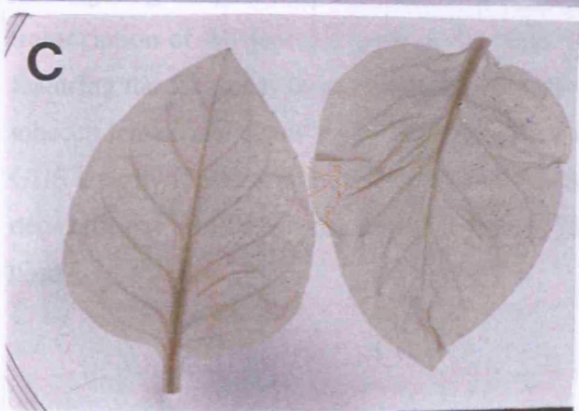
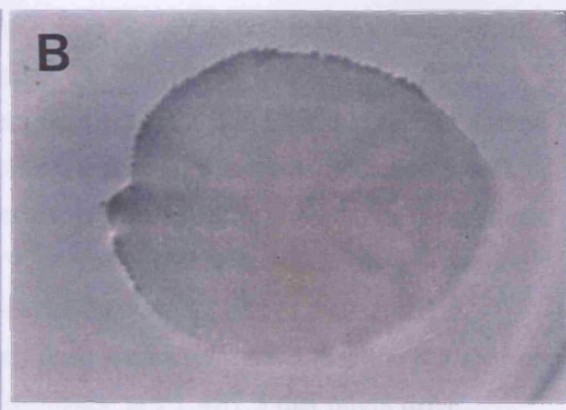
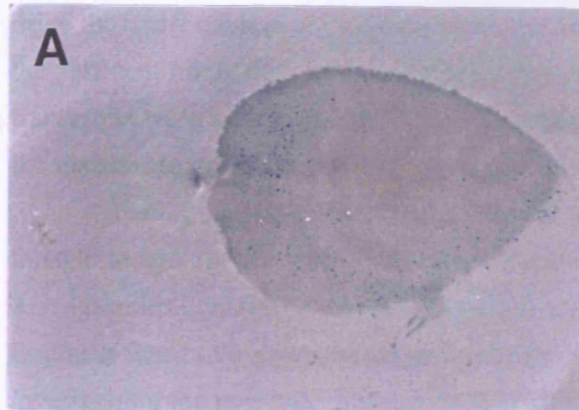
pPRP18.28-105		pPRP18.72-106		pPRP18.366		pPRP18.31-3		pPRP19-1		pPRP19-4		pPRGUS-6	
Assay	GUS	Assay	GUS	Assay	GUS	Assay	GUS	Assay	GUS	Assay	GUS	Assay	GUS
				6-1 (A)	+					6-1(A)	+	6-1 (B)	-
				6-2 (A)	+					6-2 (A)	+	6-2 (B)	
				6-3 (B)	+					6-3 (A)	+		
				6-4 (B)	+					6-4 (A)	+		
7-1 (B)		7-1 (B)						7-1 (B)	+	7-1 (B)	+	7-1 (B)	
7-2 (B)		7-2 (B)						7-2 (B)	+	7-2 (B)	+	7-2 (B)	
7-3 (B)										7-3 (B)	+		
		S-1 (A)	+										
		S-1 (A)	+										
		S-2 (A)	+										
		S-2 (A)	+										

Table 8.3 Summary of results of the transient expression assays.

The various assays performed for each construct are indicated by a reference number. The detection of GUS activity is indicated by "+". The plasmid preparation used in each assay is shown in parenthesis as "A" or "B".

Figure 8.14 Results of transient assays with constructs from the putative promoter regions. Shown are examples of tobacco leaves bombarded the various putative promoter constructs and assayed for GUS activity. Detection of *gus* activity was as described in Section 2.10.2. The plasmids used were: **A)** pPRP18.28-106; **B)** pPRP18.72-105; **C)** pPRP18.66-3; **D)** pPRP18.31-3; **E)** pPRP19-3; **F)** pPRP19-4; **G)** Transient expression assays utilizing *P. vulgaris* seeds: shown is an example of the results obtained by bombarding developing seeds with pPRP18.72-105 are shown. Prior to bombardment the seeds were previously cut open midway and the cut surface was used as a target.

The region previously described for the presence of PKT was also characterized (Section 2.4). Initially, the samples were subjected to a series of blots using anti 1 of PKT and mP1V for



The region presumed to include the promoter of *PKT2* was also characterized (Section 8.4). Initially, the anticipated ability of the segment between exon 1 of *PKT1* and mP19 to drive the expression of *PKT2* was tested (Section 8.4.1). RT-PCR was used for detection of the pertinent transcripts in the transgenic line AtEN101. The results indicated that *PKT2* transcripts were absent in AtEN101. The segment between E1.4 and E1.5 was then examined for matches to known *cis*-acting regulatory elements, and several identified (Section 8.4.2).

Testing by transient expression of the transcription activating potential of the regions thought to contain the promoters of *PKT1* and *PKT2* was performed, as described in Section 8.5. Plasmids containing genomic segments upstream from mP18 and of the E1.4E1.5 fragment fused to a downstream promoterless *gus* gene were constructed (Section 8.5.1). For investigating the possible existence of a cryptic promoter in E1.4E1.5 that might drive transcription of the reporter gene at the right T-DNA junction in AtEN101, a construct featuring the fragment in an inverted orientation was also obtained. Particle bombardment of tobacco leaves and seeds of *P. vulgaris* with the constructs was used for the transient assays of GUS activity (Section 8.5.2). For all constructs, reporter gene activity was consistently detected, but none when the negative control pPRGUS-6 including the promoterless *gus* was used.

Chapter 9

Expression of *PKT1* and *PKT2* and the application of AtEN101 as a molecular marker

9.1 Introduction

The isolation and analysis of cDNAs corresponding to transcripts mapping to the tagged region in AtEN101 revealed the presence within it of the two overlapping genes, *PKT1* and *PKT2* (Chapter 5). Relevant in assessing the functional significance of the two genes was the study of their expression patterns in the wildtype *Arabidopsis*. The relation between these and the GUS activity observed in the mutant line was also important to evaluate the usefulness and reliability of the promoter trap as a transcriptional diagnostic of tagged genes and to explore their regulation.

The study of *PKT1* and *PKT2* expression is one of the subjects of this chapter. An initial analysis to determine the distribution of *PKT1/PKT2* transcripts in the various organs of the plant was performed by Northern blot analysis (Section 9.2.1). The high structural similarity between the two genes resulting from their overlap (Chapter 5) did not allow for their discrimination by this method. For transcript-specific detection, RT-PCR was then applied to obtain information on the organ/developmental distribution of transcripts for both genes (Section 9.2.2). The expected differential expression of *PKT1* and *PKT2* in various organs was investigated following a specifically targeted strategy based on semi-quantitative RT-PCR (Section 9.2.3 and 9.2.4).

One of the most useful applications of promoter-traps is their utilization to generate molecular markers of gene expression (Section 1.3.5.4). The creation of fusion transcripts can then be exploited to study the regulatory control of the tagged gene. The response of the promoter(s) driving the expression of *gus* in AtEN101 to a number of physiological and environmental conditions was investigated to gather information that might be useful to establish the physiological significance of the tagged genes (Section 9.3).

Functional markers can also be of great benefit to the study of developmental processes (see Section 1.3.5.4). During embryogeny in AtEN101, GUS activity is observed to occur in the developing embryo in two waves. Whereas in the first period of *gus* expression the enzyme is found to be active throughout the embryo, at the cotyledonary and maturation stages GUS activity becomes progressively restricted in a polarized manner (Section 3.2.3). This characteristic of GUS expression in AtEN101 was exploited to evaluate biochemical differentiation in two embryo-defective mutants as is described in Section 9.4.

9.2 Transcriptional expression of *PKT1* and *PKT2*

9.2.1. Northern blot analysis of transcription

The distribution of *PKT1/PKT2* transcripts in different organs of C24 plants was analyzed by Northern blot hybridization. Total RNA was prepared from 9-day old seedlings and roots of seedlings grown in 1/2MS₁₀, and leaves, stems, flowers and siliques from soil grown plants. About 15 µg of each was resolved in a denaturing agarose gel (Fig. 9.1A, left) and blotted onto a Zeta-Probe[®] GT membrane (Bio-Rad). This Northern blot was hybridized with a probe prepared from the *EcoR* I insert of pCEN1-8, as described in Section 2.8.4. Autoradiographic exposure was performed for 3 weeks. In all lanes a band of hybridizing transcripts of about 1.6 kb was observed. However, the signal was considerably lower for roots. The other samples displayed bands of comparable intensity although some differences were visible with seedlings and siliques having stronger bands. The observed differences were not caused by variations in the total amounts of RNA loaded for each sample or unequal transfer in blotting, as inspection of the gel and re-hybridization of the blot to a probe prepared from a 25S *B. rapa* rRNA shows (Fig. 9.1). However, the signal was found to coincide with the 16S chloroplast rRNA band, which is far less abundant in the root sample. To verify if there was a direct relation between the two bands, washing of the blot after hybridization to the rDNA probe was performed at a lower stringency (Fig. 9.1). This had unexpected results as there was no coincidence between the observed strength of the 16S rRNA bands on the gel and the background hybridization at the same position. Nevertheless, that the detected signal obtained with the pCEN1-8 probe was not due to hybridization to the 16S rRNA is shown by careful analysis of the data. To confirm this assertion, a similar blot was prepared and hybridized to the same probes (Fig. 9.2). Comparative analysis of the gel and autoradiograms demonstrated that the 1.6 kb fragments detected with the pCEN1-8 probe are distinct from rRNA.

β-oxidation of fatty acids is a major source of energy during germination (Salisbury & Ross, 1993). It was therefore of interest to determine if *PKT1/PKT2* transcripts are induced during this period of the plant life cycle. To test this possibility a Northern blot prepared from seedling RNAs of various ages and grown in the presence or absence of sucrose was hybridized to a similar probe (Fig. 9.3). The blot was generously provided by Dr. Mark Hooks, University of Glasgow. To check for relative amounts of RNA loaded in each well the blot was subsequently stripped and hybridized with the rDNA probe. The results show a substantial induction of the *PKT* transcripts in seedlings at days 2 and 3 after germination in the absence of sucrose (Fig. 9.3). This is followed by a decrease in their relative abundance and a new increase at days 7 and 8. There is also a significant increase of *PKT* transcripts in day 2 post-germination seedlings grown in the presence of sucrose (Fig. 9.3). This induction is short-lived with the abundance stabilizing at a lower level by 3 days post-germination. There is a new increase in transcript abundance at day 8 post-germination. On the basis of a qualitative analysis the first

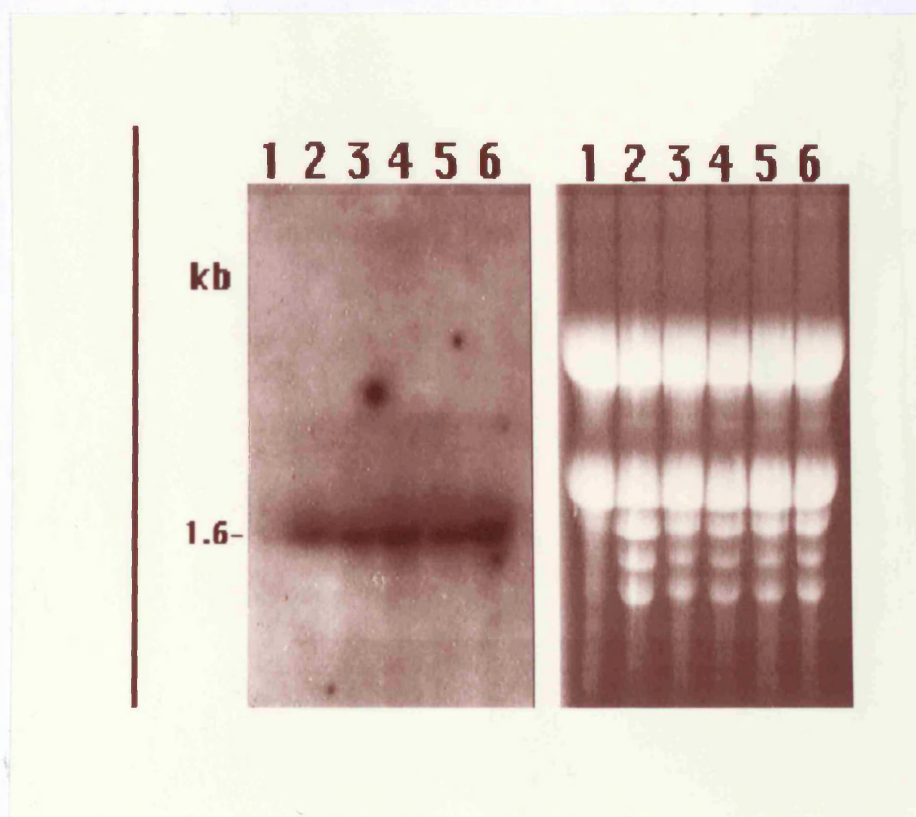
Figure 9.1 Organ distribution of *PKT1/PKT2* transcripts (I).

About 15 µg of RNA for each sample were resolved in a denaturing agarose gel. Lane 1: seedling roots (grown in 1/2MS₁₀); Lane 2: leaf; Lane 3: stem; Lane 4: siliques; Lane 5: flowers; Lane 6: 9 day-old seedlings.

A) Left: Autoradiograph of a Northern blot of the gel hybridized to the insert of pCEN1-8 (*PKT2*). Exposure lasted for 3 weeks. Right: Original denaturing gel.

B) Left: same as in A. Right: Autoradiogram of the same Northern blot hybridized to an rDNA probe prepared from pTPRH29.E3.

A)



B)

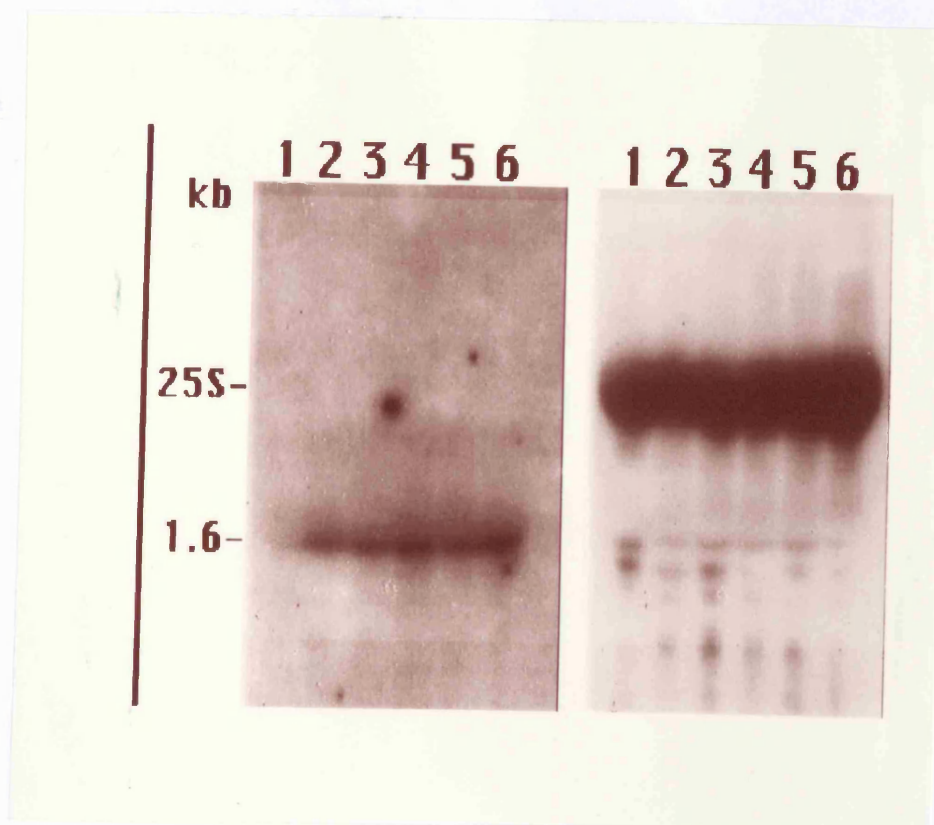


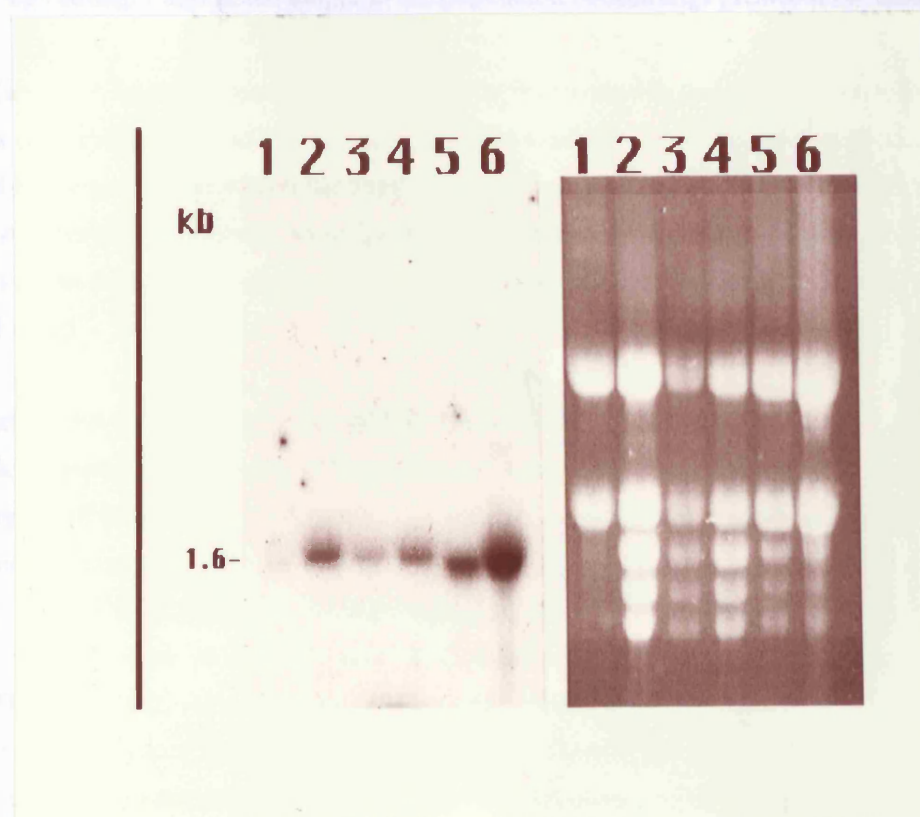
Figure 9.2 Organ distribution of *PKT1/PKT2* transcripts (II).

About 10-15 µg of RNA for each sample were resolved in a denaturing agarose gel.
Lane 1: seedling roots (grown in 1/2MS₁₀); Lane 2: leaf; Lane 3: stem; Lane 4: siliques; Lane 5: flowers; Lane 6: 5 day-old seedlings.

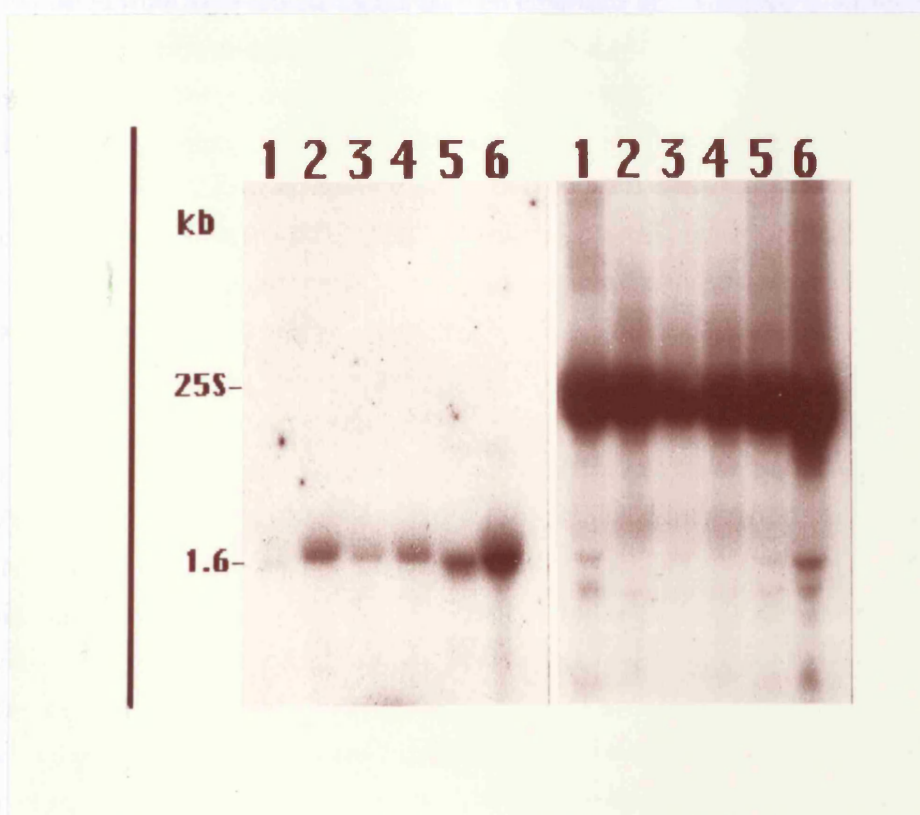
A) Left: Autoradiograph of a Northern blot of the gel hybridized to the insert of pCEN1-8 (*PKT2*). Exposure lasted for 3 weeks. Right: Original denaturing gel.

B) Left: same as in A. Right: Autoradiograph of the same Northern blot hybridized to an rDNA probe prepared from pTPH29.E3.

A)



B)



induction was stronger and lasted longer in the population of seedlings grown in the absence of sucrose.

Curiously, in some lanes, notably during the first induction period, there were present two bands of transcripts hybridizing to the probe. The bands may correspond to each of the two *PKT1* and *PKT2* genes, although other explanations are possible as discussed in Chapter 11. For example, it is possible that one of these bands corresponds to background hybridization to *PKT3* transcripts given their high similarity to those of *PKT1/PKT2* and their much higher abundance (Section 11.4.5).

9.2.2 Specific detection of *PKT1* and of *PKT2* transcripts by RT-PCR

To specifically detect the presence of transcripts for each of the *PKT1* and *PKT2* genes in the various organs of wildtype *Arabidopsis* (C24), RT-PCR was utilized. For all experiments, reverse transcription was performed on 1 µg of total RNA with a primer common to both transcripts and RT-PCR performed as described in Section 2.7.3, unless otherwise indicated. For individual amplification of *PKT1* and *PKT2* cDNAs, the transcript-specific primers prTHIO-3F and prTHIO-4F were used (Sections 5.4.2.2 and 6.3.3.1).

In an initial set of analyses the existence of *PKT* transcripts in wildtype C24 ovules at different stages of development was investigated. Ovules of up to globular stage/heart stage transition (designated here as stage 1), between late heart-stage and walking stick stage (stage 2), and embryos from cotyledonary/maturation stages (stage 3) were separately dissected. Total RNA was prepared from each sample by the method described in Section 2.6.3. Reverse transcription was primed with prTHIO-4R as was the PCR amplification of the cDNA products. Aliquots of the PCR products were resolved in an agarose gel (Fig. 9.4A, top). The results clearly showed the presence of cDNA for *PKT2* transcripts in all samples (0.5 kb; cf. Table 9.1), being less abundant in that derived from the stage 3 embryos. Also visible in the same samples are the products of amplification of the corresponding genomic fragment (1 kb). Only the latter are found in the negative control reactions. Complementary DNAs for *PKT1* transcripts (~0.4 kb) were also detected with the primer pair prTHIO-3F/prTHIO-5R in ovules of stages 1 and 2, but not in the cotyledonary stage embryos (Fig. 9.4A, bottom). To confirm the identity of the fragments, a Southern blot of the gel was prepared and hybridized with a probe derived from the insert in pCEN1-8. The hybridization pattern shows that the fragments were indeed derived from the *PKT* genes (Fig. 9.4A, right). In addition, it was possible to detect the presence of *PKT1* cDNAs in stage 3 embryos.

Similar experiments were conducted with RNA extracted from various organs. The results are exemplified in Figure 9.5A with an autoradiogram of a Southern blot of a gel of RT-PCR products of *PKT1* transcripts in silique walls, siliques, leaves, flowers and stems hybridized to a pCEN1-8-derived probe. The results clearly showed that *PKT1* transcripts were present in all samples. Carry-over contamination resulted in the weaker cDNA bands in the

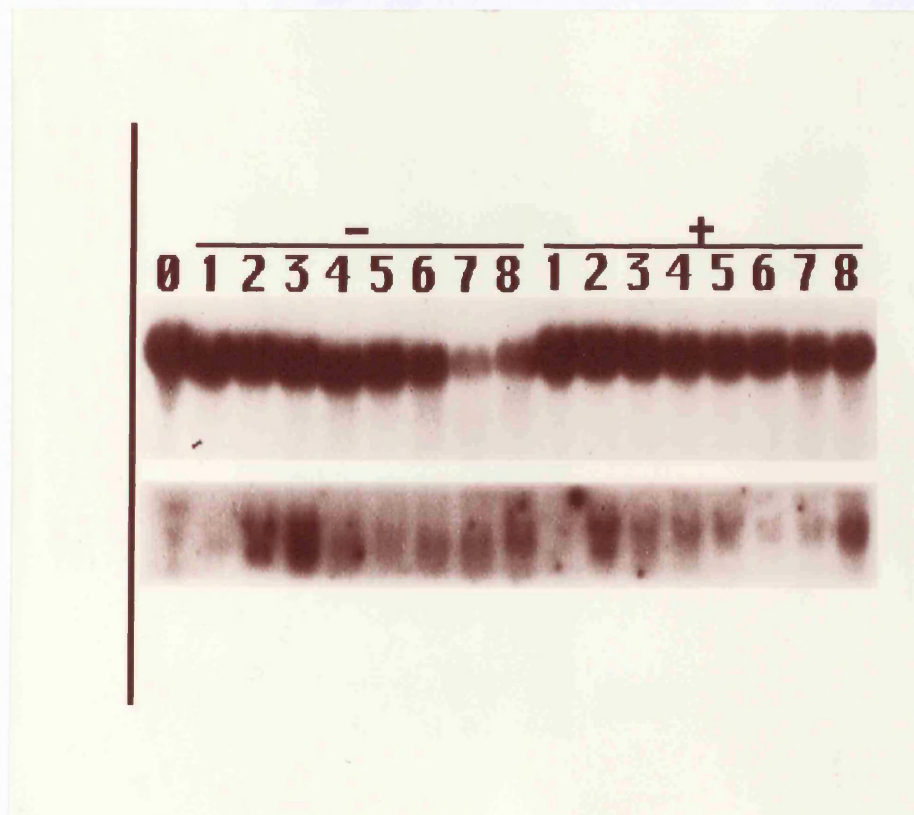


Figure 9.3 Transcription of *PKT1/PKT2* during germination.

Bottom: A Northern blot of total RNA of germinating seeds and young seedlings of different ages, and grown with or without sucrose in the media, was hybridized to a probe prepared from the insert in pCEN1-8. A corresponding autoradiogram is shown. Each lane contained 10 μ g of RNA. Lane 0: RNA from seeds imbibed for 24 h. Left, lanes 1-8: RNA from seedlings grown in the absence of sucrose (-) for 1-8 days after planting. Right, lanes 1-8: RNA from seedlings grown in the presence of sucrose (+) for 1-8 days after planting.

Top: Autoradiogram of the same Northern blot hybridized to a *B. rapa* rDNA probe prepared from plasmid pTPRH29.E3, after stripping of the first probe.

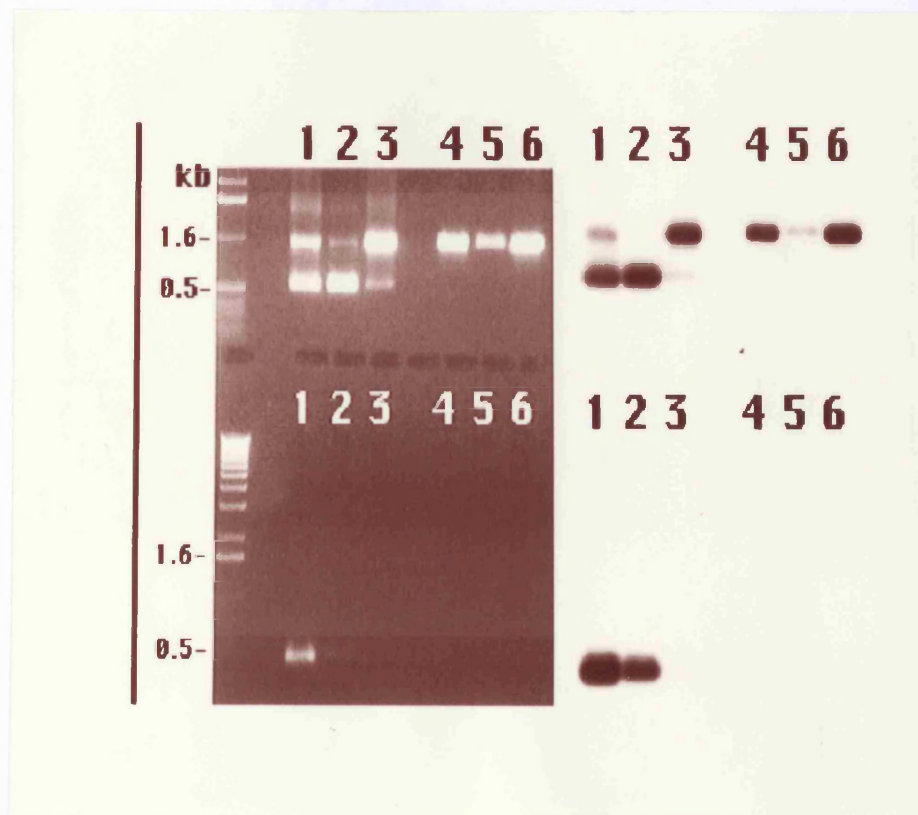


Figure 9.4 Specific detection of *PKT1* and *PKT2* transcripts in wildtype ovules.

RT-PCR was performed on RNA extracted from ovules of stage 1 (Lanes 1) and stage 2 embryos (Lanes 2), and from cotyledonary stage embryos (Lanes 3).

Bottom: Detection of *PKT2*-derived transcripts with primer pair prTHIO-3F/prTHIO-5R. Lanes 4-6: negative controls of Lanes 1-3, respectively (lacking reverse transcriptase). Left: Gel resolution of the amplified DNA; Right: Southern blot of the gel hybridized to a probe prepared from the insert in pCEN1-8.

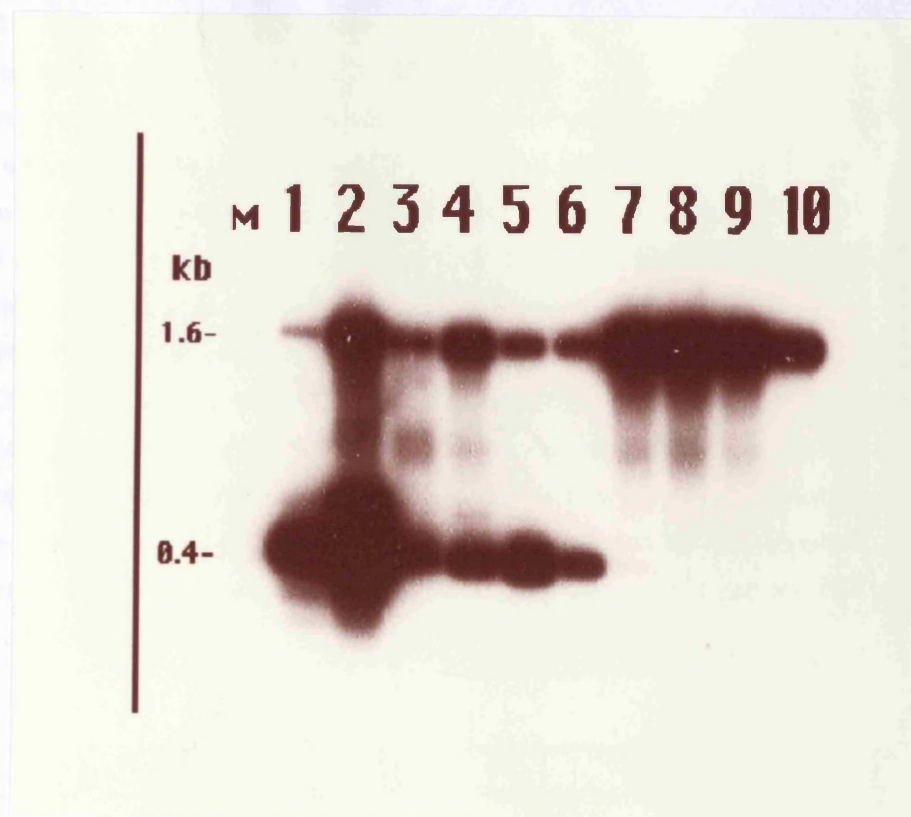
Top: Detection of *PKT1*-derived transcripts with primer pair prTHIO-4F/prTHIO-4R.

Figure 9.5 Presence of *PKT1* and *PKT2* transcripts in various organs.

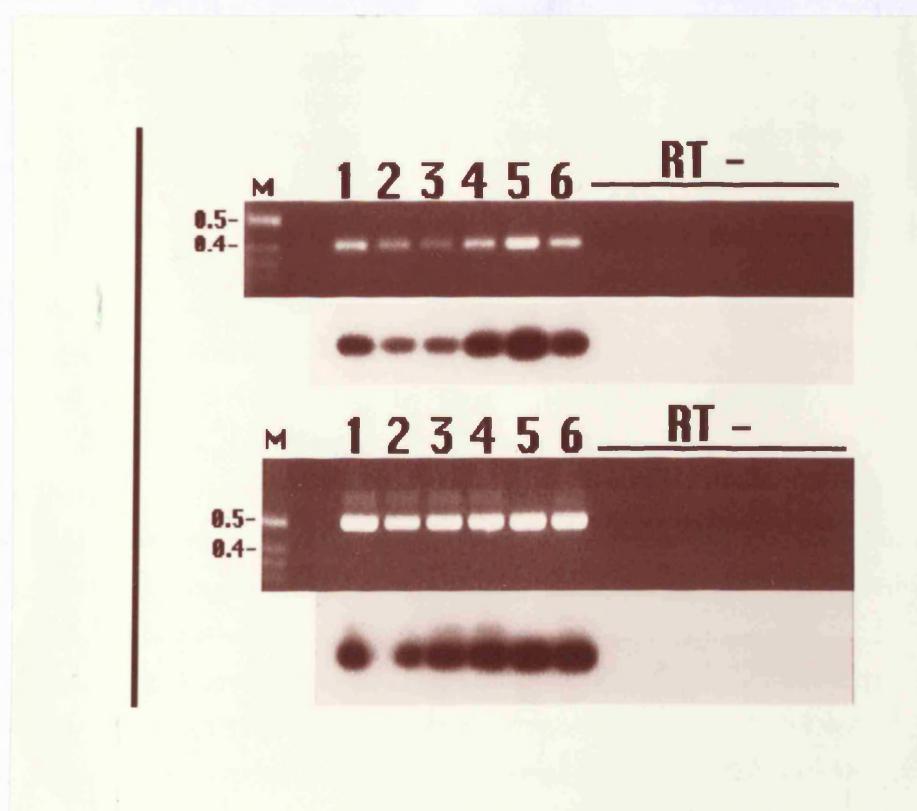
The two types of transcripts were specifically detected in samples of total RNA of different organs by RT-PCR.

- A)** Autoradiogram of a Southern corresponding to RT-PCR products for *PKT1* obtained with primers prTHIO-4R and prTHIO-3F. Lane 1: Silique walls; Lane 2: Siliques; Lane 3: Leaves; Lane 4: Flowers; Lane 5: Stems. Lanes 6-10, control reactions lacking reverse transcriptase. Reverse transcription performed with prTHIO-4R.
- B)** Detection of *PKT1* and *PKT2* transcripts following reverse transcription with primer prTHIO-5R. Shown are the gel resolved bands of amplified fragments and autoradiograms of Southern blots of the gels hybridized to a probe prepared from the insert in pCEN1-8. **Top:** Detection of *PKT1* transcripts by amplification with prTHIO-3F/prTHIO-5R. **Bottom:** Detection of *PKT2* transcripts by amplification with prTHIO-4F/prTHIO-5R. Lanes 1: Roots; Lanes 2: Leaves; Lanes 3: Stems; Lanes 4: Siliques; Lanes 5: Flowers; Lanes 6: Seedlings. The negative controls (lacking reverse transcriptase) are indicated (RT-).

A)



B)



silique wall and silique negative control reactions (Fig. 9.5A). The larger amplified fragments (~1.7 kb) correspond to genomic DNA (cf. Table 9.1)

Cleaner results were obtained when reverse transcription and amplification were performed with prTHIO-5R instead of prTHIO-4R, as this eliminates genomic DNA amplification. An example of detection of *PKT2* transcripts with amplification using prTHIO-4F and prTHIO-5R is shown in Figure 9.5B (Bottom). The results show that these transcripts were present in all organs tested including roots, leaf, stems, siliques, flowers and seedlings. *PKT1* transcripts were also detected in the same organs by RT-PCR with prTHIO-3F and prTHIO-5R (Figure 9.5B; Top). The saturation amplification with prTHIO-3F and prTHIO-4R yielded a faint band of fragments of about 0.6 kb in all of the samples except flowers (Fig. 9.5B; Bottom). This band also hybridized with the pCEN1-8 probe. Similar bands were observed in several other experiments. The nature of the band(s) is unclear except that it derives from transcripts and shows sequence similarity to the *PKT* transcripts suggesting that it may derive from an homologous gene.

PRIMERS	prTHIO-4R	prTHIO-5R
prTHIO-3F (<i>PKT1</i>)	435 nt 1734 nt	412 nt -
prTHIO-4F (<i>PKT2</i>)	543 nt 993 nt	520 nt -

Table 9.1 Length of segments between primer pairs in cDNAs and in genome. Tabulated are the sizes (in nt) of the cDNA (top) and genomic (bottom) segments that should be amplified with the primers indicated. Primer prTHIO-5R overlaps two exons and does not amplify genomic DNA.

9.2.3 Differential expression of *PKT1* and *PKT2* transcripts

Having demonstrated the existence of the two types of transcripts in the various organs their likely differential expression was investigated. To achieve this, cDNAs of both *PKT1* and *PKT2* transcripts were prepared as indicated above with primer prTHIO-5R. PCR amplification was performed in the same reaction by including both primers prTHIO-3F and prTHIO-4F. Two rounds of amplification were performed with the first consisting of 12 or 14 cycles. For the second round, 100 µl reactions were set-up and 10 µl aliquots taken at the end of a variable number of cycles. These aliquots were resolved in parallel by electrophoresis.

That the two genes are differentially regulated was shown by the dissimilar ratios of *PKT1* and *PKT2*-derived RT-PCR products obtained from distinct organs. For example, *PKT1* transcripts are proportionally more abundant in roots relative to *PKT2* transcripts than in seedlings, and least abundant in stems (Fig. 9.6).

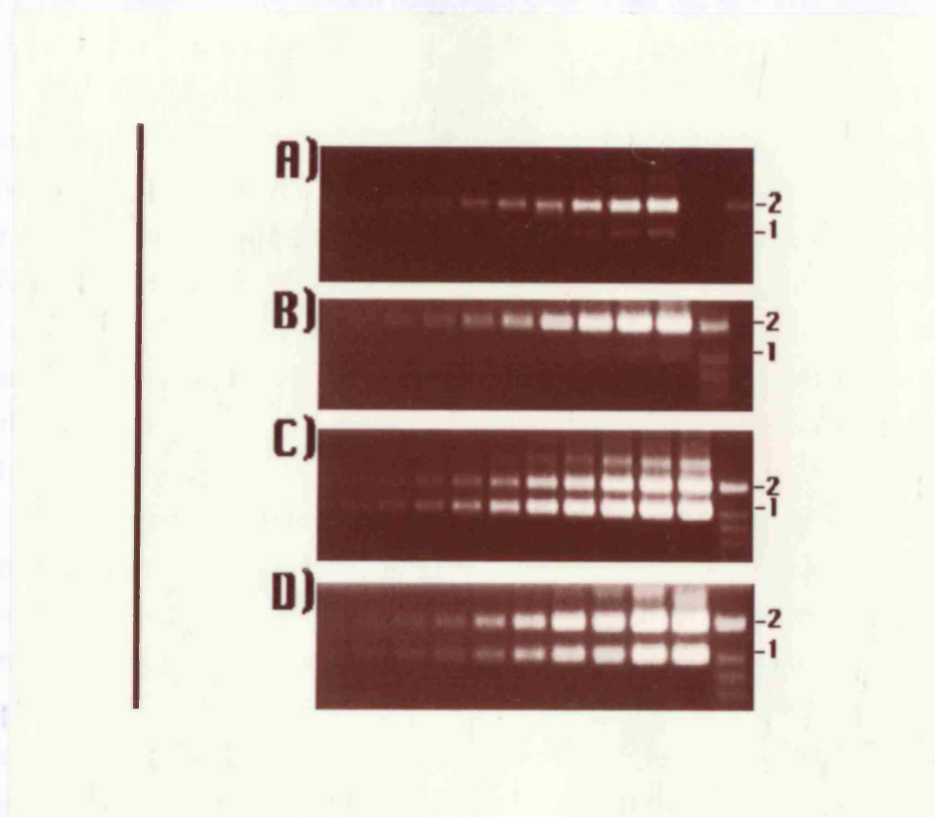


Figure 9.6 Assessing relative abundances of *PKT1* and *PKT2* transcripts.

RT-PCR for both transcripts was conducted simultaneously on total RNA. Primer prTHIO-5R was used for reverse transcription, and together with prTHIO-3F or prTHIO-4F for PCR amplification. Amplification included two rounds, the first lasting for either 12 or 14 x (94°C 25 secs, 54°C 20 secs, 72°C 35 secs). 10 μ l aliquots were taken at the end of different cycles during the second round. Gel resolution of these products is shown.

- A) Flowers;
- B) Stems;
- C) Roots;
- D) Seedlings.

9.2.4 Semi-quantification of *PKT1* and *PKT2* transcripts

With the same purpose of analyzing the differential expression of the two genes RT-PCR was used for semi-quantification of their transcripts in some organs. However, to date only a limited set of results has been obtained.

For cDNA synthesis the primer prTHIO-4R was used. In parallel, serial dilutions of linearized pCEN1-8 or pCEN1-9 plasmids were made in RT buffer supplemented with equivalent amounts of RNA sample. Both the cDNA samples and the plasmid DNA dilutions were submitted to PCR amplification. As before, the amplifications comprised two rounds with the first one consisting of 15 cycles. 10 μ l aliquots were taken at the end of various cycles for each reaction during the second round. The various aliquots were run in a gel and the relative amounts of product estimated. Comparison of the amount of products obtained with the sample and the various dilutions of the appropriate plasmid gave an indication of the abundance of the cDNA in the sample.

Preliminary experiments had established that the cDNA abundances in the various RNA samples tested were in the range of 10^4 - 10^6 molecules per μ g of RNA reverse transcribed. Therefore, this was the range of plasmid molecules used in the quantification experiments. For example, about 10^5 molecules of cDNAs from *PKT2* transcripts were estimated to be present after reverse transcription of RNA from roots and stems (Fig. 9.7D and F). They were less abundant in a leaf sample with over 10^4 but less than 10^5 molecules (Fig. 9.7G). In this same sample cDNA molecules from *PKT1* transcripts were clearly less abundant ($<10^4$ molecules/ μ g RNA; Fig 9.7B). In contrast, in samples from siliques with stage 1 embryos well over 10^5 molecules cDNAs from *PKT1* transcripts but less than 10^6 were present (Fig. 9.7A). Very similar results were obtained with RNA extracted from seeds imbibed for 24 hours (Fig. 9.7C).

9.3 Studies on the regulation of *gus* expression

The effects of several environmental and physiological stimuli on the expression of *gus* in the marker line were studied qualitatively, by the histochemical detection of its enzymic reaction. Observable changes might give indications on possible environmental control of the tagged gene(s). In essence, the analysis focused on the identification of stimuli that resulted in a strong change in the pattern of expression. Two types of stimuli were tested, including the effects of growth substances and environmental stresses. Test conditions were chosen that are known to affect the regulation of previously characterized genes. The test materials were seedlings grown hydroponically or in solid 1/2MS or 1/2MS₁₀ media under the normal conditions described in Section 2.1.4.4.

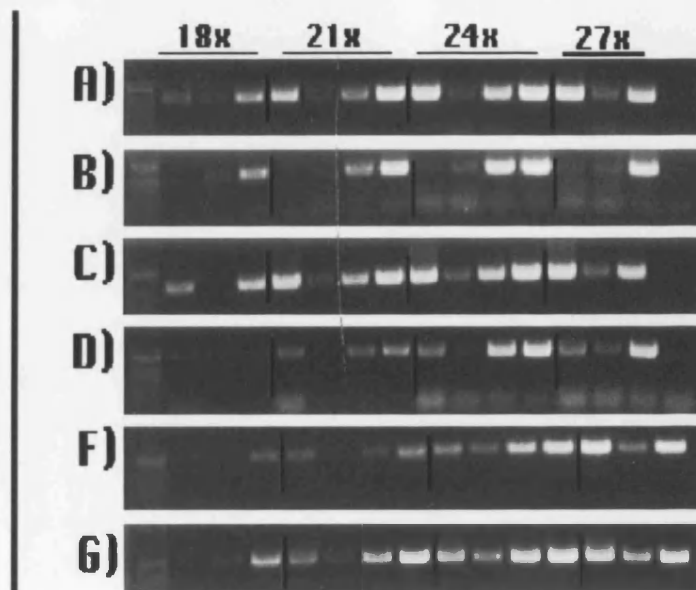


Figure 9.7 Semi-quantitation of *PKT1* and *PKT2* transcripts.

RT-PCR was performed on several samples of total RNA from different organs of wildtype *Arabidopsis*. For reverse transcription total RNA samples were reverse transcribed with primer prTHIO-4R. Parallel reactions were performed with identical aliquots of the RNA samples supplemented with known amounts of linearized pCEN1-8 or pCEN1-9 plasmids (10^4 , 10^5 and 10^6 molecules). No reverse transcriptase was added to these reaction solutions being substituted by an equivalent volume of enzyme storage buffer. For amplification all reactions relating to a same RNA source were performed in parallel. A first round of PCR 15 x (94°C 25 secs, 54°C 20 secs, 72°C 35 secs) was followed by a second round during which 10 μ l aliquots were removed from each at the end of various cycles. Shown are gels where the aliquots were run together, the cycle number at which they were sampled is indicated above. For all tests loading was: 18 x (RNA sample, standard 10^5 , 10^6); 21 x (RNA sample, 10^4 , 10^5 , 10^6); 24 x (RNA sample, 10^4 , 10^5 , 10^6); 27 x (RNA sample, 10^4 , 10^5).

A) *PKT1* transcripts in siliques (early globular to early heart stages);

B) *PKT1* transcripts in leaves;

C) *PKT1* transcripts in seeds;

D) *PKT2* transcripts in roots;

F) *PKT2* transcripts in stems;

G) *PKT2* transcripts in leaves.

9.3.1 Environmental stresses

Temperature, osmotic and wounding stresses were applied to 2-4 week old seedlings in solid 1/2MS₁₀. To study the effect of cold, seedlings were placed in a cold room (4-6°C) under dim light for 24 h, 48 h, and 72 h. Some of the plants were returned to normal growth conditions for 24 h or 48 h while others were tested immediately. No effects were observed relative to control plants. For heat treatment the plants were placed in an illuminated incubator with a day/night cycle (16h/8h) with a temperature setting of 36-42°C. Incubation periods lasted for 1 h, 3 h, and 6 h. The plants were then returned to normal growth conditions for 24 h or 48 h prior to histochemical treatment. No obvious alterations in the activity of the reporter enzyme were observed.

To test the effect of osmotic and salt stress seedlings growing on solid 1/2MS₁₀ were suspended in liquid 1/2MS₁₀ and allowed to acclimatise for 2-4 days. Then the medium was replaced with 1/2MS₁₀ supplemented with NaCl (0.25 M, 0.5 M and 1.0M) or 0.75M mannitol. The plants were submitted to the treatment for 12 h, 24 h, and 48 h after which some were incubated in 1/2MS₁₀ for 16 h prior to histochemical analysis. The distribution of GUS expression in these seedlings was similar to that of control plants, although there was wilting of the leaves in 0.5 M and 1.0 M NaCl, and poor development of leaves in 0.75 M mannitol (not shown).

Wounding was effected on leaves with a pair of tweezers. Analysis was performed 3h, 6h, 12 h, 24h, 48 h, 84 h and 1 week after inflicting the damage. Again, no obvious effects were observed (Fig. 9.8A).

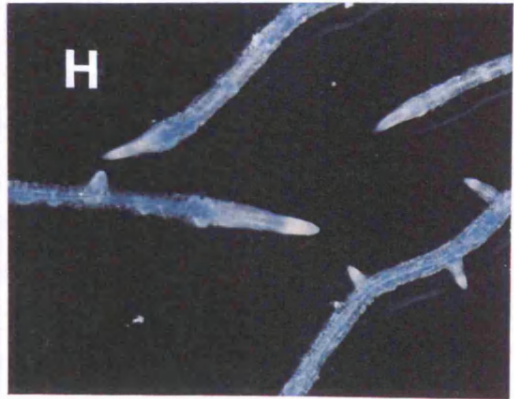
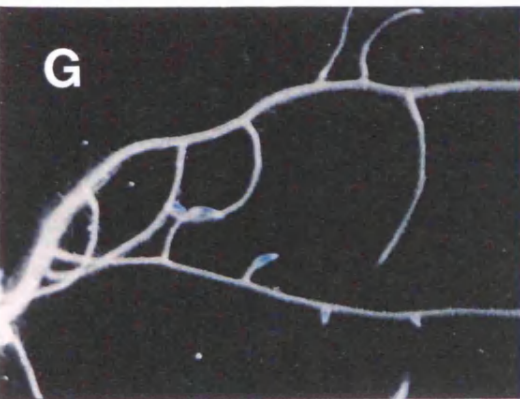
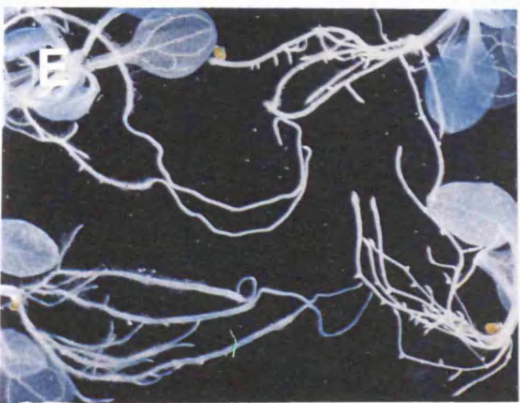
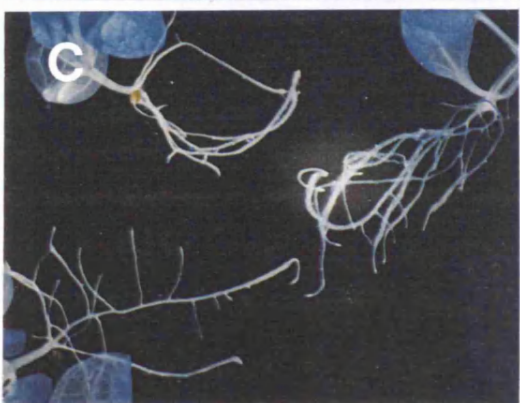
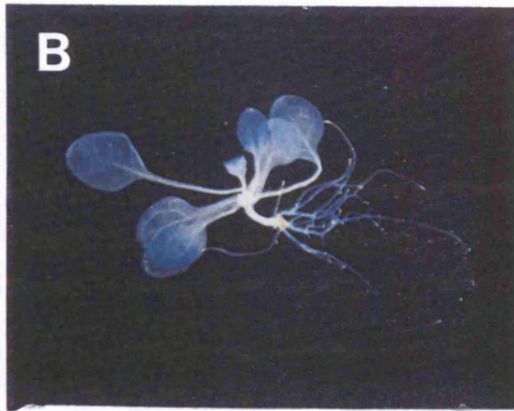
9.3.2 Effects of growth control substances

The effects of the addition of gibberellic acid (GA₃), abscisic acid, indole-3-acetic acid, and 2,4-dichlorophenoxy acetic acid to the growth medium were tested. For these experiments 2-4 week old seedlings growing in solid 1/2MS₁₀ were transferred to hydroponic cultivation in 1/2MS₁₀ or 1/2MS and allowed to acclimatise for 2-4 days. After this period the media was replaced with a fresh volume supplemented or not (controls) with a growth substance to be tested. The tested concentrations of the supplements were 10 nM, 1.0 µM, 10 µM and 100 µM. Histochemical analysis was performed on plants incubated for different periods of time in the fresh medium.

The effect of transferring the seedlings to liquid media on *gus* expression was investigated histochemically (Section 2.10.2). No obvious alterations were observed in the patterns of activity (Fig. 9.8B), except for a general decline of GUS found in plants grown in the absence of sucrose, perceptible after about 4 days of incubation (not shown).

Figure 9.8 Effects of external stimuli on the expression of GUS in AtEN101 (I)

- A) Seedling after 72 h of wounding leaves (grown in 1/2 MS₁₀).
- B) AtEN101 seedling one week after transfer to hydroponic culture from solid medium (1/2MS₁₀).
- C) Seedlings adapted to hydroponic growth in 1/2MS and exposed to various concentrations of ABA for 24 h prior to histochemical assay. Clockwise from top, right: 10 nM, 1 μ M and 10 μ M.
- D) Seedlings adapted to hydroponic growth in 1/2MS₁₀ and exposed to various concentrations of ABA for different periods. Clockwise from the Top Left: 100 μ M ABA, 36 h; 10 μ M, 36 h; 10 nM, 84 h.
- E) Seedlings adapted to hydroponic growth in 1/2MS₁₀ and exposed to various concentrations of ABA for 1 week. Counterclockwise from the Bottom Right: 100 μ M , 10 μ M, 1 μ M and 10 nM.
- F) Root system of seedling growing in 1/2MS₁₀ and exposed to 100 μ M ABA for 60 h.
- G) Root system of seedling growing in 1/2MS and exposed to 100 μ M ABA for 60 h.
- H) Root tips of seedlings grown in 1/2MS₁₀ and supplemented with 1 μ M and 10 μ M of IAA for 36 h.



Gibberellic acid (GA₃)

No noticeable change in the pattern of GUS expression was observed to be caused by growth of seedlings in GA₃-supplemented media relative to control plants, irrespective of the presence of sucrose.

Absciscic acid

ABA had a clear depressing effect on the expression of *gus* in the root systems of AtEN101 seedlings. This effect was modulated by the presence of sucrose. Thus in the absence of sucrose, after 24 h of exposure to 10 μ M ABA GUS activity had visibly decreased or been eliminated throughout most of the roots although it is retained or even increased at the tips of newly formed lateral roots (Fig. 9.8C). This effect is also seen with 100 μ M ABA (not shown), but not with the lower concentrations of the chemical (Fig. 9.8C). However, only seedlings exposed for 36 h to 100 μ M ABA, but not those in 10 μ M, showed the same alteration of GUS expression when growing in 1/2MS₁₀ (Fig. 9.8D). A concentration of 10 nM of ABA in the absence of sucrose did not have a significant effect on GUS even after 1 week of exposure (Fig. 9.8E).

The results also showed that high concentrations of ABA (100 μ M) affected the root structure by causing a relative enlargement of the younger tips. External application of ABA is known to cause alterations in root structure, that vary with concentration and result from changes in cellular growth rate, cell size and number of dividing cells (Goodwin, 1978). The observed changes were accompanied by retention of GUS activity at the tips, particularly of the younger ones, as is best seen in Figures 9.8F and 9.8G (plants exposed to 100 μ M ABA for 60 h in the presence and absence of sucrose, respectively).

Indole-3-acetic acid (IAA)

This growth regulator affected the appearance of C24 and AtEN101 seedlings and most notably their root systems, in a sucrose-dependent manner. For plants grown in 1/2MS₁₀ the effects of 1 μ M and 10 μ M IAA were visible, for example, as the initiation of numerous lateral roots after 36 h exposure (Fig. 9.8H). This resulted after 60 h in a proliferation of lateral roots for 1 μ M IAA, and the emergence of what appeared to be innumerable lateral root primordia when treatment was with 10 μ M (Fig. 9.9A). By contrast, in plants grown in 1/2MS 10 μ M IAA did not seem to cause any significant alteration in roots after 48 h (Fig. 9.9B).

GUS activity was strongest in the root primordia even prior to their emergence from underneath the epidermis (Fig. 9.8H). At higher concentrations of IAA (10 μ M) this activity was present throughout the emerging lateral roots although it was less intense closer to the tips. It is tempting to speculate that there might be a correlation between the absence of a zone lacking GUS activity and the inability of these emerging roots to develop further.

2,4-dichlorophenoxy acetic acid

The effects of 2,4-D on *gus* expression in seedlings resembled those observed with IAA, and were also influenced by the presence of sucrose in the growth media. For example, exposure to 1 μ M 2,4-D for 60 hours resulted in the appearance of numerous protrusions along the roots, particularly close to the meristematic region where they tended to coalesce together (Fig. 9.9C, 9.9D, 9.9E). GUS activity was more intense in these protrusions. As for IAA, staining was not uniform and there seemed to be a conical internal region of the root primordia where GUS activity was low, if present (Fig. 9.9D). Growth in the absence of sucrose seemed to attenuate these effects (Fig. 9.9E, 9.9F), which were absent with plants grown in the presence of 10 nM of 2,4-D. Some plants grown in 1/2MS₁₀ and exposed to 100 μ M of the 2,4-D for long periods of time (>72 h) showed a patchy distribution of GUS in their roots (Fig. 9.9G).

9.4 Expression of the marker in mutant backgrounds

The T-DNA in AtEN101 was explored as a molecular marker of embryogenic development of *gnom* and *hydra* mutants (Lindsey *et al.*, 1996; Topping *et al.*, 1997). Homozygous AtEN101 plants were crossed with heterozygous *gnom* and *hydra* plants. Seeds of the F1 generation were collected and plants grown from them. From individual F2 plants F3 seeds were separately collected. The genetic status of F2 plants relative to the T-DNA of AtEN101 was determined by plating 30-50 seeds of each plant on 1/2MS₁₀ and checking for GUS activity in young seedlings. Progeny of F2 plants shown to be homozygous for the AtEN101 trait were grown in soil. GUS activity was monitored in developing siliques and embryos and those F3 plants heterozygous for *gnom* or *hydra* identified by the presence of aberrant F4 embryos.

For both *gnom* and *hydra* heterozygous plants GUS activity was observed in silique walls and ovules until late heart stage (*e.g.* Fig. 9.10A). For the *hydra*-derived plants, although activity was also observed in developing seeds as late as torpedo stage, it was not observed in the embryos, even those homozygous for the mutation (Fig. 9.10B). Cotyledonary stage embryos were dissected and stained for GUS. Those embryos with a wildtype appearance also showed the distributions of GUS activity found in AtEN101 embryos (Fig. 9.10C). The homozygous mutant embryos also displayed staining, unevenly distributed in spite of their globular appearance (Figs. 9.10C, 9.10D).

Similar findings were obtained with the progeny of the *gnom* crosses. At the cotyledonary stage some embryos of both mutant and wildtype phenotype exhibited GUS activity but not others. In the odd shaped homozygous mutant embryos where *gus* was expressed, the distribution of the enzyme was not uniform (Figs. 9.10E-9.10H).

Figure 9.9 Effects of external stimuli on the expression of GUS in AtEN101 (II).

- A) Seedlings grown in 1/2MS₁₀ and exposed for 60 h to 1 μ M (Left) and 10 μ M (Right) IAA.
- B) Seedling grown in 1/2MS exposed for 48 h to 10 μ M IAA.
- C) Seedling grown in 1/2MS₁₀ and treated with 1 μ M 2,4-D for 60 h.
- D) Seedling grown in 1/2MS₁₀ and treated with 1 μ M 2,4-D for 60 h. Close-up of above.
- E) Seedling grown in 1/2MS and treated with 10 μ M 2,4-D for 72 h.
- F) Root of seedling grown in 1/2MS and exposed for 72 h to 100 μ M 2,4D.
- G) Seedling grown in 1/2MS₁₀ and exposed to 100 μ M 2,4-D for 84 h.

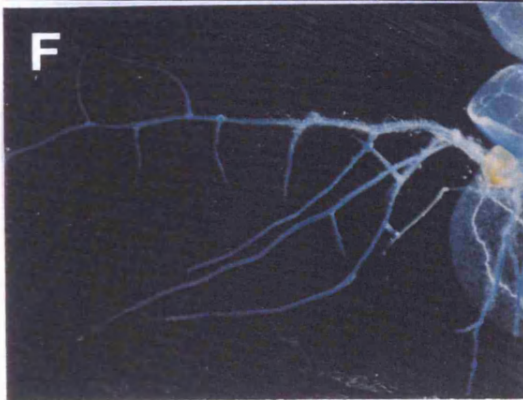
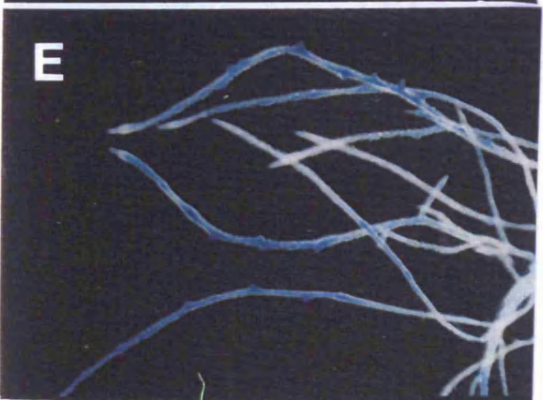
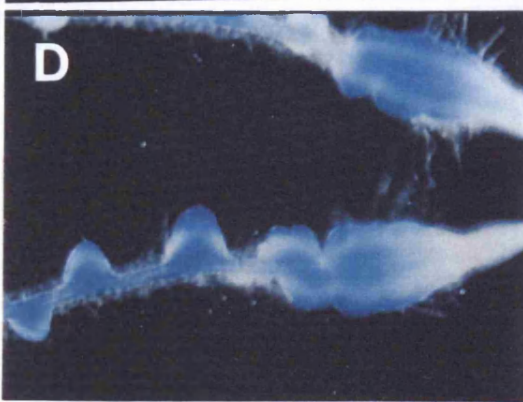
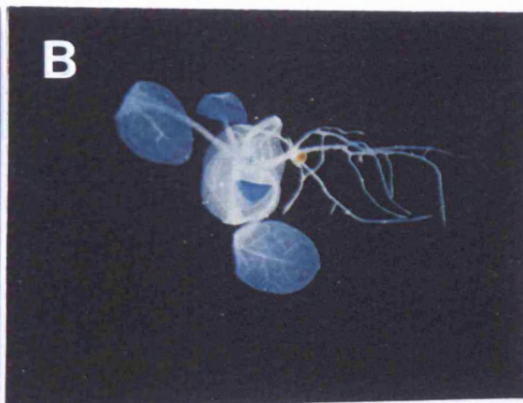
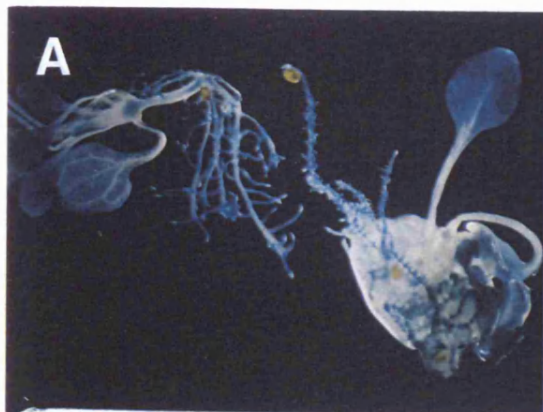
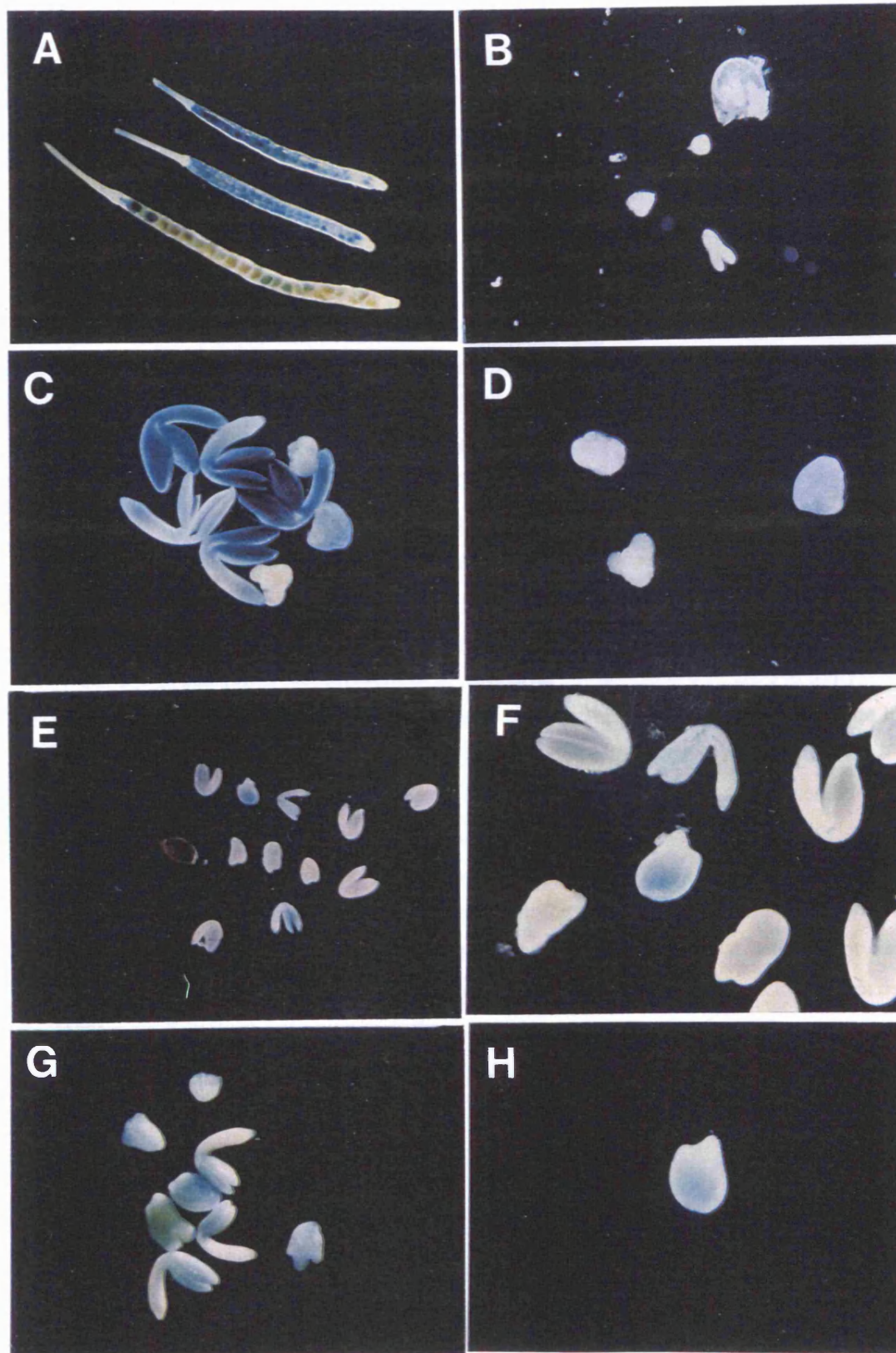


Figure 9.10 Expression of the AtEN101 reporter in *hydra* and *gnom* embryos.

- A) GUS activity in siliques of heterozygous *hydra* in AtEN101 background.
Shown are siliques with embryos at the globular and heart stage transitions, and cotyledonary stages.
- B) GUS in torpedo stage ovules.
Shown is a torpedo stage embryo (wildtype appearance) and two *hydra* embryos dissected from ovules of a same silique after histochemical analysis. Another intact ovule with an early torpedo embryo is visible.
- C) GUS in cotyledonary stage *hydra* embryos.
Cotyledonary stage embryos from a same silique were dissected and submitted to histochemical detection of GUS. Staining of the "mickey mouse" shaped mutant embryo is best seen in D.
- D) GUS in *hydra* embryos.
Another view of the same mutant embryos from a cotyledonary stage silique shown in C (above). Staining in the "ear" regions of the "mickey mouse"-shaped embryo is more clearly seen in this picture.
- E) GUS activity in *gnom* embryos (I).
A collection of cotyledonary stage embryos from a same plant is shown.
- F) GUS activity in *gnom* embryos (II).
Close-up view of some of the embryos shown in E.
- G) GUS activity in *gnom* embryos (III).
Another collection of cotyledonary stage embryos from a same plant.
- H) Polarized GUS activity in *gnom* embryo. Close-up view of one of the embryos from F.



The results of these experiments indicate that the temporal pattern of expression of the marker of AtEN101 is not affected by the *gnom* and *hydra* mutations in cotyledonary stage embryos. They also illustrate the use of this tag as a functional marker.

9.5 Summary

The expression of *PKT1* and *PKT2* at the transcriptional level in wildtype *Arabidopsis* was investigated by Northern blot analysis (Section 9.2.1) and RT-PCR (Sections 9.2.2-9.2.4). Both genes were found to be transcribed in ovules of different developmental stages, seedlings, and in roots, leaves, inflorescence stems, flowers, siliques and silique walls of mature plants. Taking advantage of the nearly identical structures of the two transcripts, differences in their relative proportions were investigated by RT-PCR (Section 9.2.3) for various RNA populations. The results demonstrated that the *PKT1* and *PKT2* are independently regulated. Data from limited semi-quantitative RT-PCR analysis corroborated this assertion and was indicative of the relative abundance of the corresponding transcripts in the RNA samples (Section 9.2.4).

The effects of environmental stresses and exposure to several growth substances on *gus* expression in seedlings of AtEN101 were investigated (Section 9.3). Of those stimuli tested only IAA, 2,4-D and ABA were found to have an effect in the pattern of GUS activity, particularly in the root system. These effects were modulated by the presence of sucrose in the growth medium. Thus IAA and 2,4-D in 1/2MS₁₀ caused alterations in roots with the appearance of numerous protuberances exhibiting GUS activity. ABA had a general depressing effect on GUS activity in roots which was attenuated in the presence of sucrose. It also caused alterations in the distribution of the enzyme in the root.

The effect of two embryogenic mutations on the expression of the functional marker of AtEN101 was investigated. The temporal pattern of expression was not affected in cotyledonary stage embryos homozygous for *hydra* and *gnom*. The polarization of GUS activity seen on AtEN101 was also observed in these embryos indicating that the activity of the reporter is regulated independently of the *HYDRA* and *GNOM* functions.

Chapter 10

A family of genes for 3-ketoacyl-CoA thiolases in *A. thaliana*

10.1 Introduction

The characterization of the T-DNA tagged region of the genome had established the existence of two overlapping genes for 3-ketoacyl-CoA thiolases in *A. thaliana* (Chapters 5 and 7). Evidence for the existence of additional isozymes had also been found, as described below. These observations argued for interesting unexpected complexities in the pathway(s) of fatty acid oxidation in *Arabidopsis*. Determining the number of genes encoding 3-ketoacyl-CoA thiolases would contribute to the unravelling of these complexities. The present chapter is concerned with contributions towards the identification and characterization of the members of this gene family.

Isozymes may result from alternative splicing of exons in the structural segment of genes. To explore the possibility that variants of the *PKT1* and *PKT2* transcripts might exist involving the use of additional exons by alternative splicing, the genomic region containing the coding portion of the genes was completely sequenced (Section 10.2). The search and identification of genes homologous to *PKT1* and *PKT2* located elsewhere in the genome of *A. thaliana* was also pursued (Section 10.3). Two different strategies were followed including the detection of genomic sequences that might harbor such genes by Southern blot hybridization analysis, under non-stringent conditions (Section 10.3.1). In addition, the possible existence of other genes for peroxisomal thiolases was also investigated by searching the database of *Arabidopsis* ESTs for corresponding transcripts. The latter strategy was powerful in that it enabled the unequivocal detection of other active genes likely encoding thiolases, and, concurrently, of eventual variants of *PKT1* and *PKT2* transcripts (Section 10.3.2). This approach resulted in the finding of one additional gene encoding a 3-ketoacyl-CoA thiolase isozyme, *PKT3*, which had also been identified by others (Newman, 1994; Cocke, 1996; Dr. Graham, personal communication).

During the writing of the present manuscript, a potential fourth gene for a 3-ketoacyl-CoA thiolase, *PKT4*, was identified in a genomic sequence of chromosome I of *A. thaliana* (Osborne *et al.*, 1997). Information on the structural and evolutionary relationships among the four (putative) *PKT* genes identified to date in *A. thaliana* was obtained by comparative analysis, as described in Section 10.4.

10.2 The structure of the *PKT1* and *PKT2* genes

10.2.1 Restriction map of pENB3

The structures of the 5' regions of the *PKT1* and *PKT2* genes were previously described (Chapter 5 and Section 8.2.3). Their characterization identified the first 5 exons and 4 introns of both genes. Together these exons corresponded to about one-third of the known lengths of the transcribed sequences. The remainder of the gene, as indicated by the results of the Southern blot analysis with the cDNA in pCEN1-8 (Section 5.3.2), was presumed to be included within the B1.3B1.4 segment of the genome. This segment had been cloned in plasmid pENB3 (Section 4.3.3). For sequencing the remainder of the *PKT1*/*PKT2* genes, primers designed previously from pCEN1-9 (Section 5.2.2) were used (see below). The sequences generated in this manner were, however, insufficient to cover the entire region of interest. Completion of the sequencing project required the generation of subclones of pENB3. As a necessary step prior to its subcloning a restriction map of the insert in pENB3 was constructed for the enzymes *Hinc* II, *Hind* III and *Pst* I. These enzymes were selected after a preliminary assay to select a suitable set to be used (data not shown). For the restriction mapping, aliquots of pENB3 DNA were digested with *Hinc* II, *Hinc* II plus *Hind* III, *Hind* III, *Hind* III plus *Pst* I, and *Pst* I. The products of the digestions were resolved in an agarose gel (Fig. 10.1A) and the sizes of the generated fragments estimated by comparison with the 1 kb DNA ladder marker (Table 10.1).

<i>Hinc</i> II	<i>Hinc</i> II + <i>Hind</i> III	<i>Hind</i> III	<i>Hind</i> III + <i>Pst</i> I	<i>Pst</i> I
# 3.0* kb	# 2.9* kb	# 4.2* kb	# 3.3* kb	# 3.3* kb
# 2.4 kb	# 1.2 kb	# 0.75 kb	# 0.80 kb	# 1.6 kb
0.65 kb	# 0.75 kb	0.75 kb	0.75 kb	# 1.1 kb
-	0.45 kb	0.37 kb	# 0.50 kb	-
-	0.35 kb	-	0.35 kb	-
-	0.30 kb	-	# 0.25 kb	-
3.4 kb	3.4 kb	3.4 kb	3.3 kb	3.3

Table 10.1 Estimated sizes of the restriction enzyme fragments of pENB3.

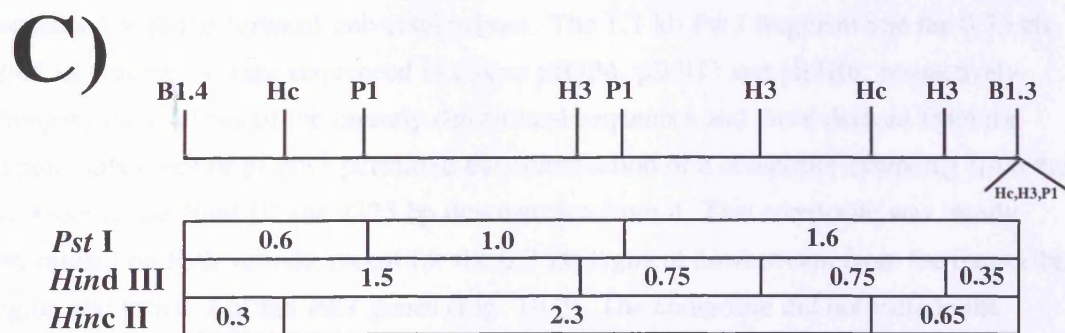
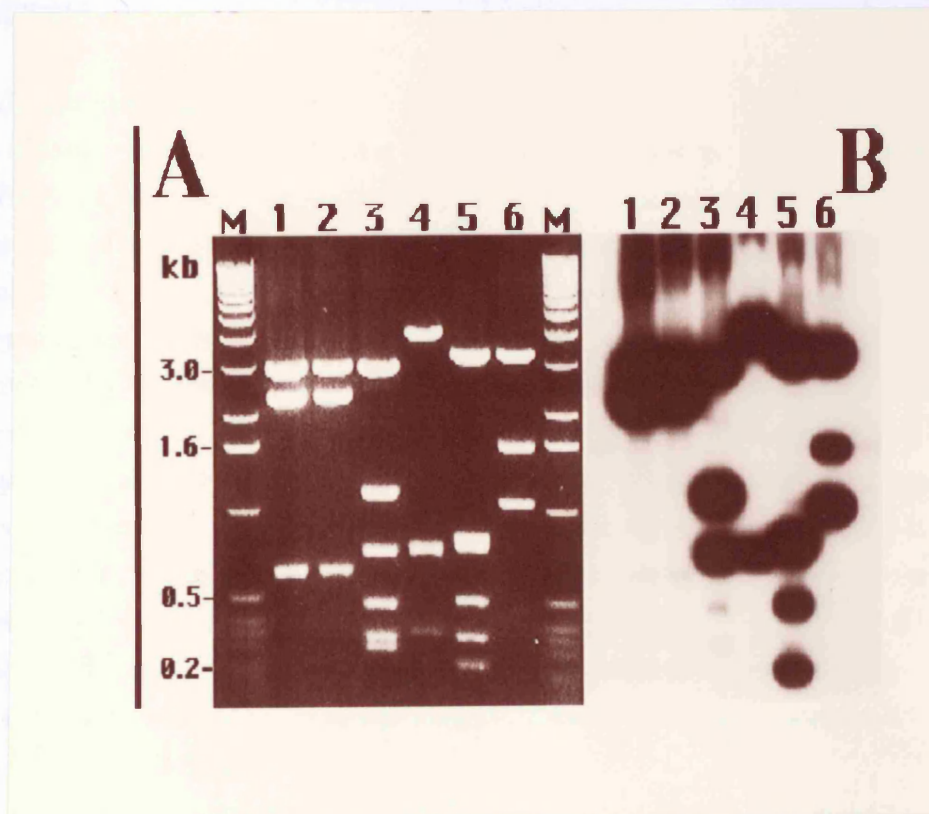
The sizes were estimated from the gel in Figure 10.1A. In the last row the estimated size of the insert is calculated for each set of fragments. Those fragments containing the 2.7 kb-long vector, pUC18, are indicated by asterisks. Fragments hybridizing to the pCEN1-9 cDNA-derived probe in the Southern blot shown in Figure 10.1B are labelled with "#".

The data gathered from the restriction enzyme analysis together with the partial sequence information of pENB3 and the known sequence of the mP18/mP19 transcripts allowed the construction of a restriction enzyme map of the plasmid (Fig. 10.1C).

Figure 10.1 Restriction enzyme mapping of the insert in the genomic subclone pENB3.

- A) Samples of plasmid DNA digested with various restriction enzymes were resolved in the agarose gel shown. Lanes M: 1 kb DNA ladder marker; Lanes 1 and 2: *Hinc* II digests; Lane 3: *Hinc* II + *Hind* III; Lane 4: *Hind* III; Lane 5: *Hind* III + *Pst* I; Lane 6: *Pst* I.
- B) Identification of the fragments on the gel containing sequences in the pCEN1-9 cDNA. An autoradiogram of a Southern blot of the gel in A) hybridized to a probe derived from the cDNA in pCEN1-9 is shown.
- C) Restriction enzyme map of the insert in pENB3. For construction of the map, the available partial sequences of pENB3 (Section 4.3.4) and the known sequences of the *PKT1* and *PKT2* transcripts were considered. The bar underneath the map shows the maximum possible extension of the *PKT1*/*PKT2* transcripts in the B1.3B1.4 region of the genome, according to the Southern hybridization data shown in B).

To determine the limits of the PKT transcription unit within the pTZ19 insert a Southern blot was prepared from the gel in Figure 10.1A and hybridized with a 32 P-labeled probe derived from the cDNA in pCFR1-9. An autoradiograph of the blot is shown in Figure 10.1B. The results clearly showed that the PKT transcription unit extended up to a maximum of about 2.5 kb in the Bgl III digest.



The *Pst* I-PvuII fragment spanning between the *Hind* III and *Pst* I sites (0.6 kb) was digested with a series of *Hinc* II in plasmid pTZ19 (section 10.1) and the resulting fragments were separated on a 1% agarose gel. The *Hinc* II fragments were then separated on a 1% agarose gel. The results are shown in Figure 10.1C. The *Hinc* II fragments were then separated on a 1% agarose gel. The results are shown in Figure 10.1C. The *Hinc* II fragments were then separated on a 1% agarose gel. The results are shown in Figure 10.1C.

To determine the limits of the *PKT* transcription units within the pENB3 insert a Southern blot was prepared from the gel in Figure 10.1A and hybridized with a [α - 32 P]dCTP-labelled probe derived from the cDNA in pCEN1-9. An autoradiogram of the hybridized blot is shown in Figure 10.1B. The results clearly showed that the *PKT* transcription units extended up to a maximum of about 2.3 kb in the B1.3B1.4 genomic fragment.

10.2.2 Sequencing of pENB3

To sequence the insert of pENB3 a dual strategy was followed. Initially the primers prTHIO-1R, -2F, -3R, -4R, prTHIOP-F and prTHIOP-R were used to directly sequence the plasmid. Of these primers only prTHIO-2F, which was later found to straddle an exon-exon boundary, failed to produce a sequence. As indicated above, completion of the sequencing project required subcloning of pENB3. Accordingly, aliquots of plasmid were digested with each of *Hinc* II, *Pst* I and *Hind* III. The fragments of interest were purified following resolution of the digests in an agarose gel. Included were the 3.3 kb and 1.1 kb *Pst* I fragments, and the two 0.75 kb *Hind* III fragments. The larger 3.3 kb *Pst* I fragment contained the original cloning vector and, for this reason, was simply circularized by self-ligation. Subcloning of the other fragments required their ligation to pBluescript KS (+) previously linearized with the appropriate restriction enzyme. The products of the ligation reactions were then used to transform *E. coli*. For quick selection of transformants of the smaller fragments, colony-PCR was used to detect the presence of inserts in the cloning vector (Section 2.5.5.4).

The sequencing strategy for pENB3 is shown in Figure 10.2. Both manual and automated sequencing were performed. Subclone pB3P1 (3.3 kb *Pst* I fragment) was sequenced with the forward universal primer. The 1.1 kb *Pst* I fragment and the 0.75 kb *Hind* III fragments were sequenced in clones pB3P4, pB3H3 and pB3H6, respectively. Comparative analysis of the directly determined sequences and those derived from the various subclones of pENB3 permitted the construction of a composite extending from the B1.4 site to the *Hind* III site 2325 bp downstream from it. This composite was mostly determined on both strands except for the 0.3 kb segment downstream from the transcribed region and intron 7 of the *PKT* genes (Fig. 10.2). The composite did not include the sequences of pB3H3 as these corresponded to the downstream-most of the 0.75 kb *Hind* III segments and it was not possible to establish their relative orientation.

The length of the sequenced genomic segment between the *Hind* III site 2325 bp from B1.4 and the point of insertion of Tn Δ XR in plasmid pTNB8-31 (Section 8.2.3) thus totalled 4938 bp (Fig. 10.3). Comparison of the genomic sequence with those of the transcribed sequences mP18/mP19 enabled the mapping of the exons of *PKT1* and *PKT2* in the region. A total of 14 exons were contabilized for each gene, all of which, except exons 1

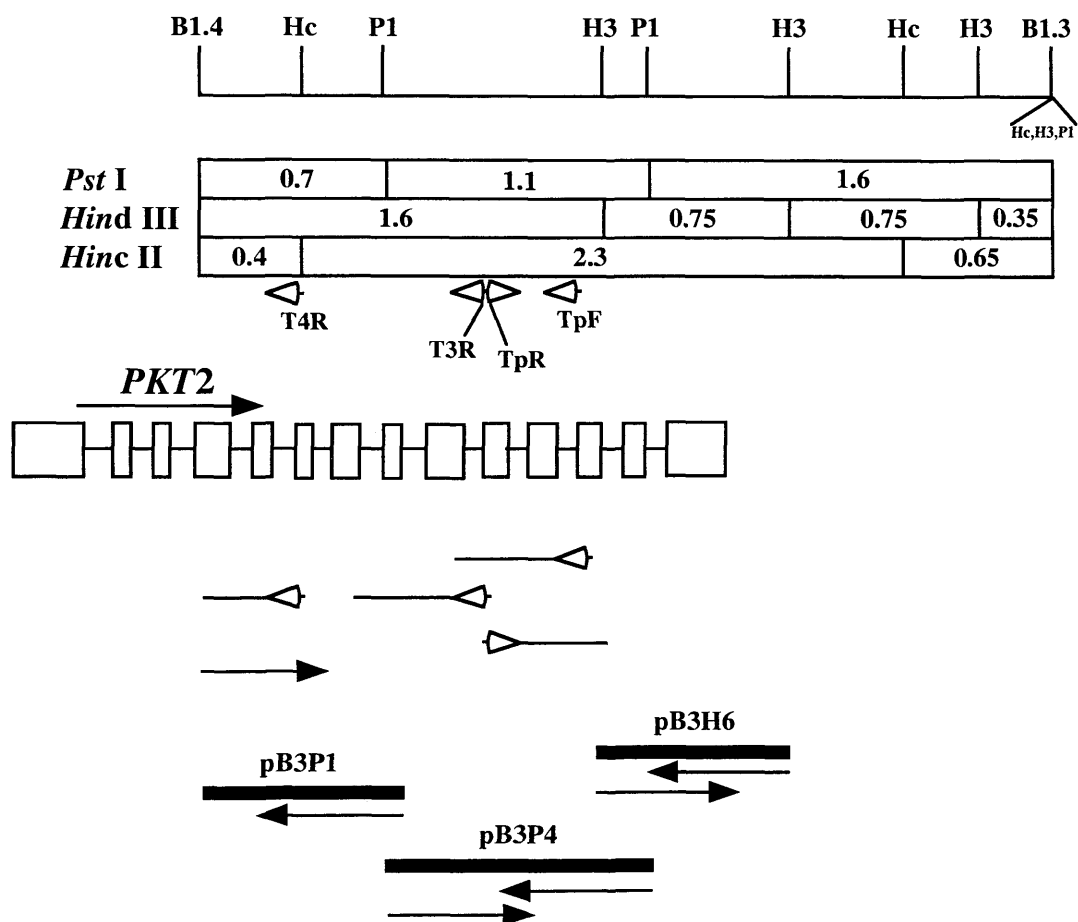


Figure 10.2 Strategy for sequencing the insert of plasmid pENB3.

The arrows indicate the length and the direction of the sequences. Their location relative to the B1.3B1.4 fragment in pNEB3 is indicated by their alignment with the restriction map for *Hind* II, *Hind* III, and *Pst* I shown. The distances between restriction sites on the map have been corrected according to the sequence data available. Arrowheads represent the sequence-specific primers used. Also shown are the fragments of the insert that were subcloned. Identified below the map are the segments corresponding to exons (open boxes) and introns (thin line) of *PKT2* as indicated by comparative analysis. The direction of transcription is shown by the thin arrow.

1 TTTTACGTTTTATGGACTTGTTGCGGCCCGTTTAGCTTTCATTTGTTATTACTAGGCCGG
61 AAACAACCCACTGGCCTTTACACCTCTATAAACCAACGCAAGTTATTTTCTTTTCGA
121 CATTGAAATTTAATATATGTTTTCATCAATCTCAAATTCATCACAATGCAGTACAAGTGA
181 ATTAAC TAAC TACACTGTAAATCTGTTTGGGTACATAAATTGTATAGCGTGAGCTACAT
241 ACAAATTTTCTTTCTTTTCGGTGTATGTATTTTTGTGTTGGTTTTTGTTTTAGGATTCTC
301 ATGGCTTTTGAATTGTTTCTAGTTATGCTTCATGTCAAAGCACAATTTCATTTAGGTTTG
361 ACAAAGCTAGTCATAGTCATGAAAAATATAGGGAGTATATTAAAGTAGGATTTTGCTCTT
421 CTCCTATTACACAACCTTGACATGCTTTTTGATAATCAATATTTATAAATAGACTATTTTA
481 TTTAAACTACTCAACAAAAAGACCCAATTTAATTGATTATTATTTAATTATGAAAATAAA
541 AATAGGACTATTATGTTTCTTTCTTTAATTTTTAAAATTCTGTTTCTAATTTGGGCTGC
601 TTAGTTGAGTAGTAATGTAGAGGACTGTTTTAATTAATTACTGAATTTTTTATTTTAAAA
661 ACAATACTATTAGTAATGTAATTTTAAAAATAAAAAATAGTATTGTTTGCATGGAATACT
721 TTGCCTCATAATGGTTTCACTTTGACTATTAGAGAATGGGAAGAAGAACTGTTTGAAGAT
781 GTTGAAC TGGTTCTTCTTCTCTTTGTCGGAGTAATCTGAATCGAAATCGACGACAACCT
841 CCATTGAAAAACATATATATATATATGTATAGATTTAAGAGAAACCTAGTAACATATTTA
901 ATCAATTCAATATTTGATTTATTGTTTTTAAATAGATAATTAACATTTTAAAATAGAAA
961 AACAAAACAGTTCAAACCTTACTTTATGTAATAAGTTAGATACTTAATAAACAAATAA
1021 AAGAAATTAGGTTTGCAAGAAATCTAACATGAATGGAAGGTAGGTAGGCATAAAAATGA
1081 AAGATAAGTAAGGAGGTTGGTAGGAAGAGTGGTGGAGGATCAAAGAGTATGGCAAAGA
1141 GTTAATTGCCAAATGCAACGTCACCTCTTTCACTCCTCATCTAACTCTCAACACTATTTT
1201 TGTATATATAATCTTGAACAATCGGTTATCTCAACATTTGCATATTGTCTACTTTTAGAT
= E5.2 =
1261 TTTGATATCTATTGGAAC TCTTACTAGCTCATT CATCCGTCGGGTAACTTTCTTGACCTC
1321 TCTCCATCTTATCTTTTACGAGTCTTAAAGACTTTACTTATCTAAGAAAATCCGTTTTAG
1381 TCTCTCAA AATTGATTTATTAGATGGTATATTCTAGTCACAACAAC TATTTTTTGGACGA
1441 GAATTTTACAAAATAATTGAAGGATACTTAACTCTTCTTACTAAACAAAACAAC TTA
= E1.5 =
1501 TCTATATCCGATATGAATTCTTGTGGGTAAAGAATACAAATCATGTATATTATCTGGATA
= E5.1 =
1561 TCCAAATTATTCACTAAAATTTTATAAATTATCTCAAATTCTTCTAAATTTGTAAATTAC
1621 TCAAATTATCAAATCACTAATACCCAAAATGAGAAGAAACATTTATTGCTTAAAAGTTA
1681 GATTCTTCCCACGTGGCCCGGTTTAGTTATAGTTGCCGACGATTAATTCCCGGTCACATT
1741 TTTAACGGCGTGATTTGTATGTGGCGTAAAACATGTCAAATGCGATATTTTCGTAACGA
== Hc ==
1800 CGTTAACTTATACAAATACGACAAGACAAGGTTTTGGCAACGAGGCCAATCAGACCAGAC
1861 TTGTGAGAACTGAGAACGATCGATCAAAAATCTCTCCCACGATAAAAAATGAGGACCCTGA
1921 AGCTCATTGTACTTAACGTCTTCCCCTTTCCAGAAAAATCAACCAAACCCATTTCAGTTT

1981 CCCCGAAAGTTTTTGTCTTTCAATCACCATTAGAATTTCAGCTGCAACACAAGTGTTC
 2041 AACAGAGAGAATCATGGAGAGAGCTATGGAAAGACAAAAGATATTGCTTCGTCATCTCAA
 2101 TCCAGTTTCTTCTTCTAATTCTTCTCTTAAACATGAACCTTCTCTTCTGTCTGTAAGTTT
 2161 AGTTGTCTTAATCCGTGAGTCTTAATTAGACATTGAAATTGTTTGTGGAAGCTGATGATC
 2221 TTTAACTCTGTTTTATTATGTGACTCACTTGTGCTTTTGTTAATTTTAGCCTGTGAATTG
 2281 TGTTTCTGAAGTTTCCCCAATGGCTGCTTTTGGAGATGACATTGTGATTGTAGCGTAAGT
 2341 ATACTCATTGAGTTTGCTTCACTTTCACCTTCTTTAGTTTTCTTTGTTTCTTGATTCTTC
 2401 TTCTTCTTCAGGGCATATCGTACCGCCATTGCAAAGCGAGACGTGGAGGTTTCAAAGAC
 2461 ACTCTTCCTGATGATCTTCTTGCTTCTGTTCTTAAGGTAAGTCATCAATTAAACCTTGTG
 2521 CATGCATGGAGTAGTAGTATCAATGATTTCTTATGTTTCATTGTTTGTCTGATTAGGCT
 2581 GTAGTGGAAGAACATCTTTGGATCCAAGTGAAGTTGGTGATATCGTTGTTGGTACCGTG
 2641 ATAGCGCCTGGTTCTCAGAGAGCAATGGAGTGTAGAGTTGCAGCTTATTTTGCTGGTTTT
 2701 CCTGGTACTAAATCTCATTCAAGAATCAAGATAAACTTAATATCTCACAAGTTGTTGATA
 2761 TGTTTCTTCATCCTTCTCTTTTACGAGCTAGACTCCGTGCCAGTTAGAACTGTCAATAGA
 2821 CAATGCTCATCAGGACTACAAGCAGTTGCTGATGTTGCTGCTTCCATTAGAGCTGGTTAT
 2881 TACGACATTGGTAAGATGCTGCTGCATCAGTTTTTTGTAGTCTCTAGCTGCTTGTATAGG
 2941 CCATGGGATTGATTCTCTCAACTTTGTGAATTGCATAGGTATTGGTGCTGGAGTGGAAATC
 3001 AATGTCAACTGATCATATTCTTGAGGCGGCTTTCATGGCTCTAATCCAAGAGTAAGCTA
 3061 TCACAATGAGTGTATAAATTTACTCAACAAGAAGAGTTTTGTGATGATCTTGTAAGTCT
 3121 ACATGGGAATAGGCACAGGATTTCCCAAAAGCCCGTGATTGTTTGCTTCCAATGGGAATT
 3181 ACTTCTGAAAACGTTGCAGAAAGGTTCGGTGTCACAAGAGAAGAGCAAGATATGGCTGCG
 3241 GTTAGTTAGTTATTCTCTTCACTTTTTTCTCTCTAGTTCATATAAGTACTGATTCAGTA
 3301 ATAATCAGGTGGAGTCTCACAAACGCGCTGCAGCTGCAATCGCGTCTGGTAAACTCAAGG
 3361 ATGAAATCATTCCTGTTGCTACTAAGGTAAAACAAGAACCTCCCAAACCTCCTTTAAAGC
 3421 CCTTTTGGTAGAGCTTCTCAGTAAATGATATTGTTCTTGACAGATTGTGGACCCTGAGA
 3481 CTAAAGCAGAGAAGGCAATCGTCGTATCTGTTGATGACGGTGTACGTCCAAACTCAAACA
 3541 TGGCAGATTTGGCAAAGCTGAAGACTGTCTTTAAACAGAACGGTTCACCACAGCTGGTT
 3601 AGAAACAACCTCTCTTAACTATCACTGGCATAAGTGAGATTCATAGATTGAAGCATTTAAG
 3661 TATATGTGTTTTGTGTAGGCAATGCTAGTCAGATCAGTGATGGTGCTGGAGCAGTACTGC
 3721 TAATGAAGAGAAGTTTGGCTATGAAGAAGGGACTTCCCATTCCTGGAGTATTCAGGTAAA
 3781 ATGAATATACAACATTTTGCCAGTTCAAGAAACCGAGCAAATCTGATTCCGGAACAATGC
 3841 AGGAGTTTTGCTGTTACTGGTGTGGAACCATCTGTAATGGGTATTGGTCCAGCTGTTGCC

and 2 of *PKT1*, were separated by short introns (67-117 bp). The complete description of the exons and introns in the known structural portion of the two genes is given in Table 10.2.

The principal conclusion of this analysis was the demonstration that no additional exons were present within the limits of the transcription units as indicated by the small size of the introns. The data also permitted the calculation of the minimal length of the primary transcripts of *PKT1* and *PKT2* as 3394 bp and 2771 bp, respectively.

Element	Location	Length	Element	Location	Length
Exon 1 (1)	1241-1303	63 bp	Intron 7	3241-3308	68 bp
Intron 1 (1)	1304-2269	966 bp	Exon 8	3309-3386	78 bp
Exon 1 (2)	1864-2152	289 bp	Intron 8	3387-3464	78 bp
Intron 1 (2)	2153-2269	117 bp	Exon 9	3465-3597	133 bp
Exon 2	2270-2334	65 bp	Intron 9	3598-3678	81 bp
Intron 2	2335-2411	77 bp	Exon 10	3679-3775	97 bp
Exon 3	2412-2496	85 bp	Intron 10	3776-3842	67 bp
Intron 3	2497-2577	81 bp	Exon 11	3843-3966	124 bp
Exon 4	2578-2704	127 bp	Intron 11	3967-4049	83 bp
Intron 4	2705-2791	87 bp	Exon 12	4050-4158	109 bp
Exon 5	2792-2890	99 bp	Intron 12	4159-4240	82 bp
Intron 5	2891-2978	88 bp	Exon 13	4241-4324	84 bp
Exon 6	2979-3052	74 bp	Intron 13	4325-4408	84 bp
Intron 6	3053-3132	80 bp	Exon 14	4409-4634	226 bp
Exon 7	3133-3240	108 bp	-	-	-

Table 10.2 List of the exons and introns of *PKT1* and *PKT2*.

Indicated are the locations of the various elements in the genomic sequence shown in Figure 10.3, together with their lengths. The distinct first exons and introns of *PKT1* and *PKT2* are indicated by the corresponding gene number within parenthesis.

10.3 Genes encoding isozymes of *PKT1* and *PKT2*

10.3.1 Searching for additional genes by Southern blot analysis

Southern blot hybridization is a useful tool for the identification of DNA segments with sequence similarity to the hybridizing probe. The lower limit of similarity required for detection is dependent on several factors, but notably on the stringency of the hybridization conditions and subsequent washes. These features were used to attempt the identification of

other genomic segments with limited similarity to *PKT1/PKT2* that might correspond to additional genes for thiolases.

Samples of DNA from *A. thaliana* cv. Columbia were digested with the enzymes *Bam*H I, *Eco*R I, *Eco*R V, *Hinc* II, *Kpn* I and *Xba* I. The choice of enzymes was based on the known restriction enzyme patterns of *PKT2*. The digested DNAs were resolved in a 0.7% agarose gel. A Southern blot was obtained from the gel and later hybridized with a probe prepared from purified cDNA insert in pCEN1-9. The probe was labelled with [α - 32 P]dCTP by the oligonucleotide random priming method. The hybridization was performed under the usual conditions (Section 2.8.3) except that the temperature was kept at 55°C.

The temperature of hybridization was chosen taking in consideration that, under identical reaction conditions, 65°C proved to be a suitable temperature for homologous hybridization of the cDNA in pCEN1-8 to the genomic DNA (Section 5.3.2). The melting temperature of hybrids between heterologous molecules drops by about 0.5-1.4°C per percentil of dissimilarity, with an average value of 1°C (Anderson & Young, 1985). Therefore, the temperature used, 55°C, was adequate for detection of genes sharing an overall exon structure and with about 85% or greater sequence similarity to *PKT2*.

An autoradiogram of the hybridized blot is shown in Figure 10.4. The sizes of the fragments visible in the autoradiogram are listed in Table 10.3. In it, most of the bands corresponded to fragments of the expected size for the *PKT1/PKT2* genomic region or were otherwise present when a similar hybridization was conducted at high stringency, as described in Section 5.3.2. However, the sizes of other fragments were clearly incompatible with the restriction patterns of the *PKT1/PKT2* genomic region. This was the case for the 3.6 kb, 2.8 kb and 6.5 kb *Eco*R I fragments. Similarly, the 6.6 kb *Eco*R V and the 4.7 kb *Xba* I fragments could not correspond to the *PKT1/PKT2* region of the *Arabidopsis* genome.

<i>Bam</i> H I (lane 1)	<i>Eco</i> R I (lane 2)	<i>Eco</i> R V (lane 3)	<i>Hinc</i> II (lane 4)	<i>Kpn</i> I (lane 5)	<i>Xba</i> I (lane 6)
> 12 kb	6.5 kb	6.6 kb	# 2.3 kb	> 10 kb*	8.9 kb
# 3.4 kb	# 4.6 kb	# 4.9 kb	# 1.2 kb	8.5 kb	8.3 kb
-	3.6 kb	# 1.1 kb	-	6.9 kb	7.0 kb
-	2.8 kb	-	-	4.7 kb	4.7 kb

Table 10.3 Sizes of fragments hybridizing to the *PKT2* cDNA in pCEN1-9.

The fragments hybridizing under medium stringency conditions as seen in the Southern blot shown in Figure 10.4 are listed. Fragments found to also hybridize at high stringency (Section 5.3.2) are indicated by "#". The asterisk for the larger than 10 kb *Kpn* I fragments indicates that these are clustered in more than one band.

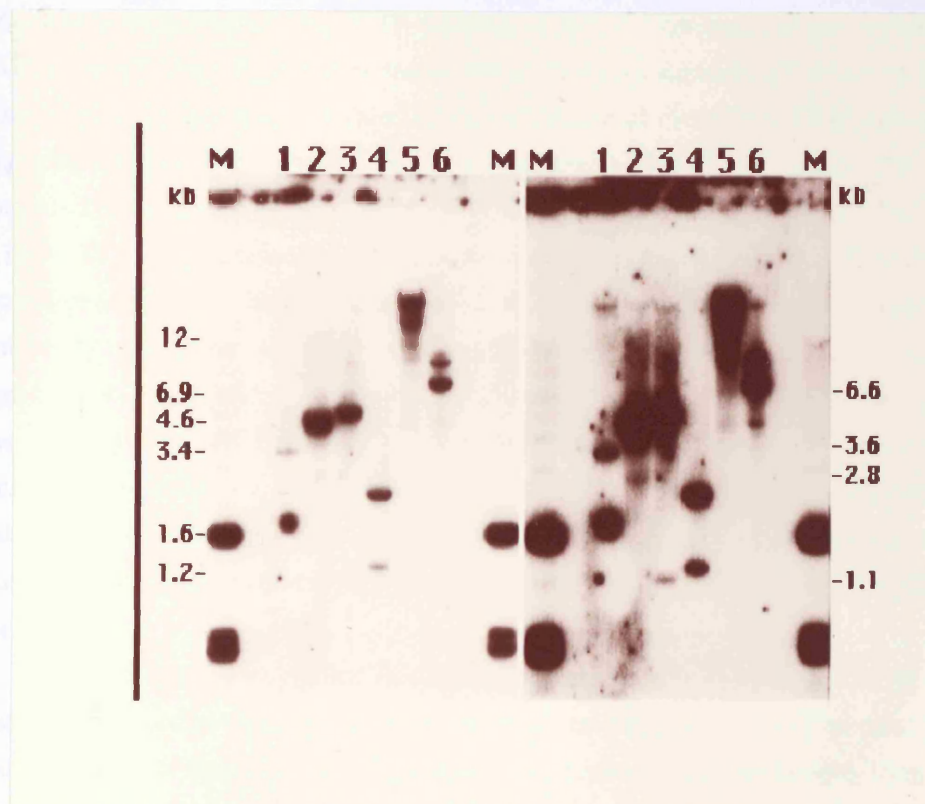


Figure 10.4 Southern blot detection of genomic regions with similarity to the *PKT1/PKT2* genes.

A Southern blot of a gel resolving digest of genomic DNA of *A. thaliana* cv. Columbia was hybridized with a probe prepared from the insert in the cDNA plasmid pCEN1-9. Hybridization was under mid-stringency conditions as detailed in the text. Shown are two autoradiograms of the blot exposed for different periods in order to more clearly see the various hybridizing fragments. Lanes M: 1 kb DNA ladder marker; Lane 1: *Bam*H I digest; Lane 2: *Eco*R I; Lane 3: *Eco*R V; Lane 4: *Hinc* II; Lane 5: *Kpn* I; Lane 6: *Xba* I. The estimated sizes of several hybridizing fragments are indicated.

10.3.2 Identification of ESTs of *PKT* genes

The database of EST sequences at the *A. thaliana* Database (Flanders *et al.*, 1998), AtDB, was searched for sequences potentially encoding proteins similar to PKT1 and PKT2 with the BLAST algorithm (Altschul *et al.*, 1990). For the search parameters used, 27 EST sequences were detected with ORFs showing significant similarity to the putative PKT1 and PKT2 proteins (Fig. 10.5). Only one of these, GenBank accession T75842, clearly derived from a *PKT1* transcript as indicated by the alignment of its ORF to PKT1 and confirmed by alignment of their corresponding nucleotide sequences. The 5' end of this EST sequence was found to coincide with position 27 in mP18 (cf. Section 5.5; Appendix 6). The genic origin of this EST was confirmed by sequence analysis of the corresponding cDNA clone, accession 147C20T7, obtained from the *Arabidopsis* Biological Resource Center (Ohio State University, Columbus, Ohio), as described in Section 5.5. Sequencing of the terminal regions of its insert with the universal primers showed that the cDNA derived from a *PKT1* transcript. Another EST recently added to the database, GenBank accession AI100014, was found to derive from the 3' end of either a *PKT1* or *PKT2* transcript as shown by the protein and nucleotide sequence alignments (Appendix 7). The polyadenylation site on both EST transcripts coincided with the furthest downstream site mapped in the *PKT1/PKT2* transcripts.

Careful comparative analysis of the remaining ESTs showed that 21 of the cDNA clones corresponded to transcripts of a distinct gene for peroxisomal 3-ketoacyl-CoA thiolase. Some of these ESTs had previously been identified as originating from transcripts of a thiolase gene by Newman *et al.* (1994), Cocke *et al.* (1996) and Dr. I. Graham (personal communication). The gene, here designated *PKT3*, was first identified by Rounsley *et al.* (1997) in a genomic sequence of chromosome II. The protein it encodes was recently shown to be a peroxisomal 3-ketoacyl-CoA thiolase by Hayashi *et al.* (1998) who named the gene *PED1*.

The other four ESTs detected in the search (GeneBank accession numbers H36507, T45052, T76117 and AA404882) were very similar to each other, as is shown in the multiple alignment in Appendix 8, and seem to derive from the same gene. They were, however, distinct from those of the identified *PKT* genes. All four ESTs originated from the 3' regions of cDNAs and the C-termini of their encoded ORFs showed similarity to those of thiolases. Figure 10.6 shows an alignment of the PKT gene products with a translation of a consensus sequence derived from the ESTs where the resemblance is evident. Because the sequence data available was limited to the 3' end of the cDNAs, the identity of the gene(s) represented in these ESTs is yet to be established (see Chapter 11).

Results of BLAST query of EST sequences with PKT2

Sequences producing High-scoring Segment Pairs:	Reading Frame	High Score	Smallest Sum	
			Probability P(N)	N
GenBank AI099670 AI099670 33823 CD4-13 Arabidopsis tha...	-1	461	5.8e-44	1
GenBank ATTS3781 Z35342 A. thaliana transcribed sequen...	+1	380	3.6e-35	1
GenBank R89952 R89952 16307 Lambda-PRL2 Arabidopsis th...	+3	358	4.0e-33	1
GenBank ATTS4063 Z37598 A. thaliana transcribed sequen...	+3	309	9.9e-28	1
GenBank T75842 T75842 10620 Lambda-PRL2 Arabidopsis th...	+2	235	1.2e-26	2
GenBank W43093 W43093 22474 Lambda-PRL2 Arabidopsis th...	+1	284	7.2e-25	1
GenBank T45659 T45659 8922 Lambda-PRL2 Arabidopsis tha...	+1	276	1.0e-23	1
GenBank N37839 N37839 19066 Lambda-PRL2 Arabidopsis th...	+3	223	3.0e-23	2
GenBank ATTS3836 Z35397 A. thaliana transcribed sequen...	+1	259	7.2e-22	1
GenBank T04395 T04395 442 Lambda-PRL2 Arabidopsis thal...	+2	216	1.6e-21	2
GenBank ATTS4548 Z47692 A. thaliana transcribed sequen...	+1	255	1.9e-21	1
GenBank N38511 N38511 19738 Lambda-PRL2 Arabidopsis th...	+2	230	6.7e-19	1
GenBank H36507 H36507 14175 Lambda-PRL2 Arabidopsis th...	+2	229	8.5e-19	1
GenBank H37590 H37590 15719 Lambda-PRL2 Arabidopsis th...	+1	220	1.1e-17	1
GenBank AI100014 AI100014 34167 Lambda-PRL2 Arabidopsi...	+2	215	3.8e-17	1
GenBank T45052 T45052 8315 Lambda-PRL2 Arabidopsis tha...	+2	214	4.6e-17	1
GenBank T20943 T20943 2951 Lambda-PRL2 Arabidopsis tha...	+2	161	1.4e-16	2
GenBank ATTS3424 Z34664 A. thaliana transcribed sequen...	-3	151	1.4e-16	2
GenBank N65908 N65908 20948 Lambda-PRL2 Arabidopsis th...	+3	206	3.7e-16	1
GenBank T76117 T76117 10895 Lambda-PRL2 Arabidopsis th...	+1	205	4.4e-16	1
GenBank ATTS2754 Z32626 A. thaliana transcribed sequen...	+1	203	9.6e-16	1
GenBank R65341 R65341 13845 Lambda-PRL2 Arabidopsis th...	+1	184	1.2e-13	1
GenBank T12940 T12940 1006 Lambda-PRL1 Arabidopsis tha...	+1	182	1.2e-13	1
GenBank AA712266 AA712266 31994 Lambda-PRL2 Arabidopsi...	+3	182	1.5e-13	1
GenBank AA404882 AA404882 28266 Lambda-PRL2 Arabidopsi...	+1	145	1.2e-09	1
GenBank ATTS0545 Z18059 A. thaliana transcribed sequen...	+1	95	1.6e-07	2

Figure 10.5 ESTs encoding ORFs with similarity to PKT1 and PKT2.

The conceptual translations of the *Arabidopsis* EST sequences deposited at the AtDB database (Flanders *et al.*, 1998) were searched for similarity to the putative proteins PKT1 and PKT2 (February 1999). The search was conducted on-line with a BLAST program (Altschul *et al.*, 1990) that searches a protein query sequence against nucleotide sequences translated in all reading frames, TBLASTN. The comparison matrix used was BLOSUM62 (Henikoff & Henikoff, 1992). Underlined are the two ESTs derived from *PKT1* or *PKT2* transcripts (see text).

		*		20	*		40																																				
PKT2-COOH	:	AM	KK	GL	PIL	GV	FR	SFA	VTG	VE	PS	VM	GIG	PA	VA	IP	PA	AT	K	L	A	G	:	41																			
PKT3-COOH	:	AM	QK	GL	PVL	GV	FR	TFA	AV	G	VD	PA	IM	GIG	PA	VA	IP	PA	AV	K	A	A	G	:	41																		
PKT4-COOH	:	AM	QK	GL	PIL	GV	FR	TFS	AV	G	VD	PA	IM	GV	G	PA	VA	IP	PA	AV	K	A	A	G	:	41																	
TEST	:	AL	Q	L	G	L	L	V	L	A	K	I	K	G	Y	G	D	A	A	Q	E	P	E	F	F	T	T	A	P	A	L	A	I	P	K	A	I	A	H	A	G	:	41

		*		60	*		80																																				
PKT2-COOH	:	LN	V	S	D	I	D	L	F	E	I	N	E	A	F	A	S	Q	Y	V	Y	S	C	K	K	L	E	L	D	M	E	K	V	N	V	N	G	G	A	I	A	:	82
PKT3-COOH	:	LE	L	D	D	I	D	L	F	E	I	N	E	A	F	A	S	Q	F	V	Y	C	R	N	K	L	G	L	D	P	E	K	I	N	V	N	G	G	A	M	A	:	82
PKT4-COOH	:	LE	L	N	D	V	D	L	F	E	I	N	E	A	F	A	S	Q	F	V	Y	C	R	N	K	L	G	L	D	A	E	K	I	N	V	N	G	G	A	I	A	:	82
TEST	:	LE	S	S	Q	V	D	Y	Y	E	I	N	E	A	F	A	V	V	A	L	A	N	Q	K	L	L	G	I	A	P	E	K	V	N	V	N	G	G	A	V	S	:	82

		*		100	*		120																																					
PKT2-COOH	:	I	G	H	P	L	G	A	T	G	A	R	C	V	A	T	L	L	H	E	M	K	R	R	G	K	D	C	R	F	G	V	I	S	M	C	I	G	T	G	M	G	:	123
PKT3-COOH	:	I	G	H	P	L	G	A	T	G	A	R	C	V	A	T	L	L	H	E	M	K	R	R	G	K	D	C	R	F	G	V	V	S	M	C	I	G	T	G	M	G	:	123
PKT4-COOH	:	I	G	H	P	L	G	A	T	G	A	R	C	V	A	T	L	L	H	E	M	K	R	R	G	K	D	C	R	F	G	V	V	S	M	C	I	G	S	G	M	G	:	123
TEST	:	L	G	H	P	L	G	C	S	G	A	R	I	L	I	T	L	L	G	I	L	K	K	R	--	N	G	K	Y	G	V	G	G	V	C	N	G	G	G	A	:	121		

		*		140	*																												
PKT2-COOH	:	AA	A	V	F	E	R	G	D	S	V	D	N	L	S	N	A	R	V	A	N	G	D	S	H	-----	:	148					
PKT3-COOH	:	AA	A	V	F	E	R	G	D	G	V	D	E	L	R	N	A	R	K	V	E	A	Q	G	L	L	S	K	D	A	R	:	154
PKT4-COOH	:	AA	A	V	F	E	R	G	G	G	V	D	E	L	C	D	V	R	K	V	-----	:	143										
TEST	:	S	A	L	V	L	E	-----	:	127																							

Figure 10.6 Alignment of the consensus ORF with the PKT sequences.

The consensus derived from the multiple alignment of the ESTs with the GeneBank accession numbers H36507, T45052, T76117 and AA404882 (see Appendix 7) was translated. The resulting ORF was subjected to multiple alignment with PILEUP to the C-terminal regions of the PKT2 and PKT3 sequences. The alignment parameters were gap penalty=3.0; Gap extension penalty=0.1. Positions where equivalent residues are conserved in all three sequences are shadowed black. Grey shadowing marks positions where equivalent residues are conserved in two of the sequences. Detection of conserved residues and shadowing was performed with GeneDoc (Nicholas *et al.*, 1997).


```

      *      20      *      40      *      60
PKT3 : MEKATERQRVLLRHLP--SSSSSHNYEAS-LSASACLAGDSAAQCRTSLYGDDVVIVAARHTPL : 62
PKT4 : MEKATERQRIILRHLP--SSSS----DAS-LSASACLSKDSAAQ----YGDDVVIVAARHTAL : 54
PKT2 : MERAMERQKILIRHINPVSSNS--LKHEPSLLSPVNCVSEVSP----MAAFGDDIVIVAAYRTAI : 61
      <<----      Exon 1      ---->><<----      Exon 2      ---->><<----

      *      80      *      100      *      120      *
PKT3 : CKSKRGNFKDTYPDLLAPVLR---ALIEKTNLNPSEVGDIVVGTVLAPGSQRASECRMAAFYAG : 124
PKT4 : CKAKRGSEKDTFPDELLASVLRVCLALIEKTNVNPSEVGDIVVGTVLGPGSQRASECRMAAFYAG : 119
PKT2 : CKARRGGFKDTLPDDLASVLK---AVVERTSLDPSEVGDIVVGTVIAPGSQRASECRVAAYFAG : 123
      Exon 3      ---->><<----      Exon 4      ----

      140      *      160      *      180      *
PKT3 : FPETVAVRTVNRQCSSGLQAVADVAAAIKAGFYDI-----GIGAGLESMTTNFM---AWEGSV : 179
PKT4 : FPETVFIRTVNRQCSSGLQAVADVAAAIKAGFYDIGKIVTIFGIGAGLESMTTNER---GWKGSV : 181
PKT2 : FPDSEVVRTVNRQCSSGLQAVADVAAASIRAGYYDI-----GIGAGVESMTDHIIPGGGFHGS- : 180
      >><<----      Exon 5      ---->><<----      Exon 6      --

      200      *      220      *      240      *      260
PKT3 : NPANVKKFAQAQNCCLPMGVTSENVAAQRFVGSRQEQDQAAVDSHRKAATAAGKFKDEIIPVKTG : 244
PKT4 : NPNVKKFEQAHCNCLPMGITSENVAAHNFVNSREEQDQAAVDSHRKAASATASGKFKDEIIPVKTG : 246
PKT2 : NPRAQDFPKARDCLPMGITSENVAAERFVGTREEQDMAAVESHKRAAATAAGKFKDEIIPVATK : 245
      -->><<----      Exon 7      ---->><<----      Exon 8      ---->>

      *      280      *      300      *      320
PKT3 : LVDPKTGDEKPITVSVDGIRPTITTLASLGKLPVFKKDGTTTA-----GNSSQVSDGAGAVF : 302
PKT4 : IVDPKTGDEKPITVSVDGIRPNTTISGLAKLPVFKEDGTTTAGTFPRLIGNSSQLSDGAGAVL : 311
PKT2 : IVDPETKAEKATVSVSDGVRPNNSMADLAKLKVFKONGSTTA-----GNASQISDGAGAVL : 303
      <<----      Exon 9      ---->><<----      E

      *      340      *      360      *      380      *
PKT3 : LMKPSVANQKGLPVLGVFRTFAAVGVDPAIMGIGPAVAIPAAVKAAGLELDIDLFEINEAFASQ : 367
PKT4 : LMRNIVANQKGLPILGVFRTFSAVGVDPAIMGVGPAVAIPAAVKAAGLELNDVDLFEINEAFASQ : 376
PKT2 : LMKRSLAMQKGLPILGVRSFAVTGVESVMGIGPAVAIPAATKLALNVSDIDLFEINEAFASQ : 368
      xon 10      ---->><<----      Exon 11      ---->><<----

      400      *      420      *      440      *
PKT3 : FVYCRNKLGLDPEKINVNGGAMAIGHPLGATGARC VATLLHEMKRRGKDCRFVVSVMCIGTGMGA : 432
PKT4 : FVYCRNKLGLDAEKINVNGGAIAIGHPLGATGARC VATLLHEMKRRGKDCRFVVSVMCIGSGMGA : 441
PKT2 : VVYSCKKLELDMEKINVNGGAIAIGHPLGATGARC VATLLHEMKRRGKDCRFVVISMCIGTGMGA : 433
      Exon 12      ---->><<----      Exon 13      ---->><<----

      460      *      480
PKT3 : AAVFERGCGVDELRLNARKVEAQGLLSKDAR : 462
PKT4 : AAVFERGCGVDELCDVRKVH----- : 461
PKT2 : AAVFERGDSVINTSNARVANGDSH----- : 457
      Exon 14      ---->>

```

Figure 10.7 Multiple alignment of the (putative) PKT2, PKT3 and PKT4 proteins.

The alignment was performed with the program PILEUP of the GCG package (Devereux *et al.*, 1984). A gap penalty of 5.0 and a gap extension penalty of 0.3 were applied for the alignment, and the comparison matrix was the normalized table of Dayhoff (Gribskov & Burgess, 1986). Positions where equivalent residues are conserved in all three sequences are shadowed black. Grey shadowing marks positions where equivalent residues are conserved in two of the sequences. Detection of conserved residues and shadowing was performed with GeneDoc (Nicholas *et al.*, 1997).

10.3.3 A fourth, putative gene for peroxisomal 3-ketoacyl-CoA thiolase

Within the international initiative to sequence the entire genome of *A. thaliana* currently in progress, an additional putative gene for a peroxisomal 3-ketoacyl-CoA thiolase was identified in the BAC clone T1G11 of chromosome I by Osborne *et al.* (1997). The genomic sequence of the region hypothetically encoding the gene, the arrangement of exons and introns and the putative sequence of the translated products, as deduced by Osborne *et al.* (1997), are deposited at the GenBank under the accession number AC002376. The gene candidate is here designated *PKT4*. With the purpose of confirming that *PKT4* was a genuine gene the *A. thaliana* EST database was searched for transcripts originating from it, but none were found.

10.4 Comparative analysis of the *PKT* genes

10.4.1 The exon/intron structure of the *PKT* genes

The study of the organization of the different elements in the (putative) primary transcripts of the *PKT* genes may contribute to the understanding of their evolutionary relationships. For this reason, a multiple alignment of *PKT2*, *PKT3* and *PKT4* was produced and the exon boundaries in each of the sequences were identified. These boundaries were established for *PKT2* (and *PKT1*) in this work (see Table 10.2), for *PKT3* by Hayashi *et al.* (1998), and for *PKT4* by Osborne *et al.* (1997). As is clear from the alignment, the four genes have the same number of protein-coding exons and of introns. In addition they seem to share the same exon boundaries along the length of their (putative) protein products (Fig. 10.7). The alignment also showed that the small insertions observed in the sequence of the hypothetical *PKT4* (Osborne *et al.*, 1997) relative to the other proteins are located at the junctions between its exon pairs 3/4, 5/6, and 10/11. Examination of the genomic sequences of the relevant exon/intron junctions of *PKT4* revealed the presence of other potential 5' splice site signals for each exon pair that, if used, would result in the elimination of the small insertions in the alignment (Fig. 10.8). Taking in consideration these corrections of the *PKT4* suggested structure, *PKT4* becomes more similar to the other three proteins.

10.4.2 Relatedness of the *PKT* genes

Pairwise comparison of the three proteins indicated that *PKT1*/*PKT2* are more distantly related to *PKT3* and to *PKT4* than the latter are to each other (Table 10.4). The differences reflected subjacent levels of relatedness of the corresponding (putative) transcripts, as is shown in Table 10.5 for the pairwise alignments of their protein-coding segments. The corrected sequence of *PKT4* (Section 10.4.1) and its corresponding portion of the transcript were used for these analysis.

Exon 3/Intron 3

....TCTGTATTGAGAGGTATGTCTTGTTCAG....
....S V L R V C L

Exon 5/Intron 5

....TATGACATTGGTAAGATAGTAACTATCTTTGGTTACA....
....Y D I G K I V T I F G

Exon 10/Intron 10

....ACAACTGCTGGTACTTTCCCCAGATTAAATTGGTCATTTTCAGA....
....T T A G T F P R L I G

Figure 10.8 Suggested corrections for the splicing model of the *PKT4* transcript.

Shown are the alternative 5' splice sites that are suggested to be used during maturation of the putative *PKT4* primary transcripts. The dinucleotide GT consensus at the proposed splice sites are double underlined. The dinucleotide consensus at the sites suggested by Osborne *et al.* (1997) are also underlined. The amino acid residues eliminated from the PKT4 sequence by the corrections are shown in bold typeface.

	PKT3	PKT4
PKT2	0.860 / 0.713	0.862 / 0.722
PKT3	-	0.946 / 0.823
PKT4	-	-

Table 10.4 Identity and similarity levels between the *Arabidopsis* thiolases.

The identity (bold typeface) and similarity levels between the putative proteins were derived from pairwise alignments performed with GAP (Devereux *et al.*, 1984), implementing gap penalties of 3.0 and gap extension penalties of 0.1.

	<i>PKT3</i>	<i>PKT4</i>
<i>PKT2</i>	0.714	0.693
<i>PKT3</i>	-	0.840
<i>PKT4</i>	-	-

Table 10.5 Identity between *PKT* transcripts (coding regions).

Identity levels were calculated for pairwise alignments of the protein-coding portions of (putative) transcripts of the *PKT* genes. The alignments were performed with GAP (Devereux *et al.*, 1984). Gap penalties were of 5.0 and gap extension penalties were of 0.3.

The above comparisons were, however, derived from the sequences corresponding to the complete lengths of the (putative) proteins. A more accurate estimate of their relatedness requires the elimination of the weighing uncertainties associated with the gaps observed in their multiple alignment (cf. Fig. 10.7). For this reason the regions contributing for the presence of gaps in the multiple alignment were removed from the sequences prior to a new set of pairwise alignments. The segments removed corresponded in PKT2 to the 46 N-terminal residues, the sequence PGGGFHGS at position 177, and the C-terminal NGDSH residues. The two remaining segments for each sequence were joined into composites each totalling 395 residues. These composites were submitted to divergency analysis with the NEWDIVERGE program (Devereux *et al.*, 1984). This program estimates the percentage of residue substitutions between pairs of sequences. Using NEWDIVERGE with the Kimura protein distance correction, it was estimated that the relative distance of PKT2 to either PKT3 or PKT4 was about three times the distance between the latter two (Table 10.6).

As expected, the rate of divergency is much higher for the introns of the genes. Although the genes share the same exon structure the variation in sequence and in size of their introns is considerable. This can be seen in the multiple alignment in Appendix 9.

Inspection of this alignment also clearly showed that the rate of divergency is not identical for all introns.

	PKT3	PKT4
PKT2	29.42	31.30
PKT3	-	11.50
PKT4	-	-

Table 10.6 Estimated divergency of the (putative) peroxisomal thiolases.

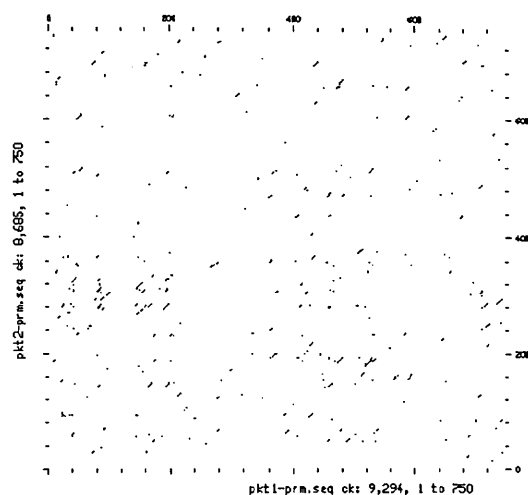
The 395 residues-long ungapped composites of the protein sequences (see text) were analyzed with the program NEWDIVERGE (Devereux *et al.*, 1984) implementing the Kimura protein distance correction. The numbers indicate estimated number of substitution per 100 residues.

For example, introns 1 varied in size from 117 bp in *PKT2*, to 238 bp in *PKT4* and 412 bp in *PKT3*. However, sequence similarity was easily recognizable, particularly between *PKT3* and *PKT4*. More dissimilar were the introns 12 which displayed very limited sequence conservation and varied in size from 65 bp in *PKT4* to 82 bp in *PKT2* and 287 bp in *PKT3*. In contrast, intron 3 was almost identical in size in all genes being shorter 1 bp in *PKT4* relative to the others, and its sequence was reasonably well conserved with 25% of the positions being identical in all three sequences and 50% identical in two of them.

In an attempt to identify a possible conserved structure in the promoter/ 5' Untranslated Region (5' UTR) sequences of the genes, the 750 bp-segments immediately upstream from the first AUG codon were aligned with the multiple alignment tool PILEUP of the GCG package (Devereux *et al.*, 1984). These segments include known 5' UTRs of 93 bp for *PKT1*, 183 bp for *PKT2* and 108 bp for *PKT3*. This failed to show any conserved motifs at similar relative positions (data not shown). To detect possible conservation of features between pairs of sequences all possible pairwise combinations were analyzed with the program COMPARE of the GCG package. For windows of a specified length this program detects similarities above a minimum threshold irrespective of their relative positions in the sequences being analyzed. The results of the analysis plotted with DOTPLOT are shown in Figure 10.9. In all cases the observed similarities were restricted to very short segments. To better evaluate the possible significance of these short stretches of sequence similarity a program that detects localized segments of sequence similarity in multiple sequences, MACAW (Schuler *et al.*, 1991), was used. A number of short blocks of similarity were found but none was statistically significant (data not shown). In conclusion, overall the similarity between the various genes upstream from the translated regions was very limited.

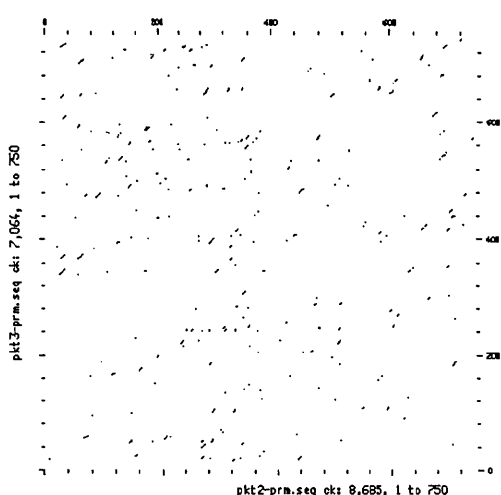
Figure 10.9 Pairwise comparisons of the 5' non-translated regions of the *PKT* genes. The 750 bp upstream from the protein coding portions of the (putative) genes were compared pairwise the COMPARE (Devereux *et al.*, 1984). Plotted with DOTPLOT are the results. For the analysis windows of 12 residues and a stringency of 9 hits were used.

OUTPUT of: pkt1-prm_10211.pnt Density: 854.35 November 3, 1998 13:33
 CORRAGE Window 12 Scrlngeny: 9.0 Points: 701



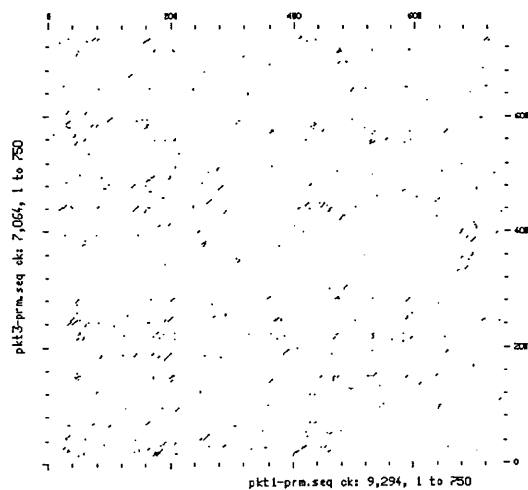
pkt1 vs. pkt2

OUTPUT of: pkt2-prm_3336.pnt Density: 854.35 November 3, 1998 17:34
 CORRAGE Window 12 Scrlngeny: 9.0 Points: 550



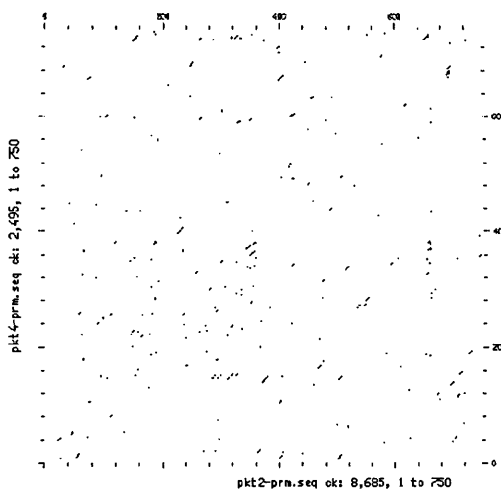
pkt2 vs. pkt3

OUTPUT of: pkt1-prm_20173.pnt Density: 854.35 November 3, 1998 16:13
 CORRAGE Window 12 Scrlngeny: 9.0 Points: 827



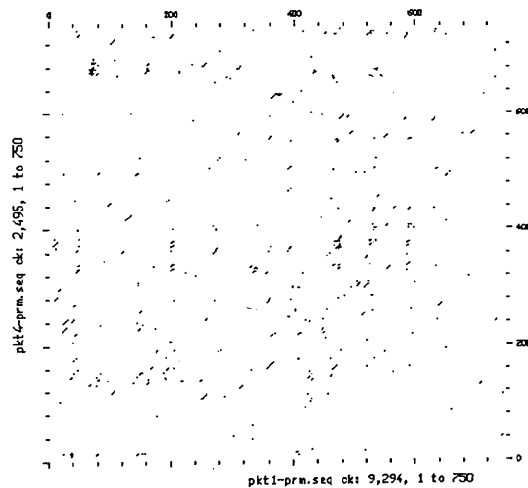
pkt1 vs. pkt3

OUTPUT of: pkt2-prm_20143.pnt Density: 854.35 November 3, 1998 17:00
 CORRAGE Window 12 Scrlngeny: 9.0 Points: 498



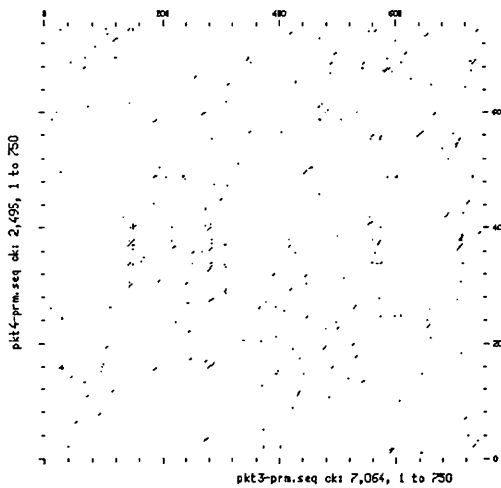
pkt2 vs. pkt4

OUTPUT of: pkt1-prm_22150.pnt Density: 854.35 November 3, 1998 16:34
 CORRAGE Window 12 Scrlngeny: 9.0 Points: 757



pkt1 vs. pkt4

OUTPUT of: pkt3-prm_27120.pnt Density: 854.35 November 3, 1998 17:07
 CORRAGE Window 12 Scrlngeny: 9.0 Points: 590



pkt3 vs. pkt4

Chapter 11

Discussion

11.1 Overview of the results

The general objective of the research presented in this thesis was to contribute to the study and assessment of the T-DNA interposon p*Agus*Bin19 as a promoter trap and as an insertional mutagen, and the exploitation of both these aspects to the study of a particular developmental phenomenon. The work centered on two case studies including of a transgenic line (AtEN101) lacking an obvious phenotype but expressing the reporter gene in a temporally and spatially controlled manner, and on the identification of mutations affecting morphogenesis that might have been caused by T-DNA insertion. The main progress achieved in each area is briefly summarized below.

11.1.1 The transgenic line AtEN101

Molecular characterization of the tagged genomic region

- Through isolation and analysis of the tagged region in the wildtype and the transgenic line, and of cDNAs mapping to it, the T-DNA was found to be inserted in the first long intron of a gene, *PKT1*. The first exon of a second gene, *PKT2*, was located closely downstream. *PKT1* and *PKT2* overlapped and shared the same exon/intron structures, except for the first exon/intron pair.
- A single T-DNA locus was present in AtEN101. The insertion had a complex structure and left border regions were found at both junctions with the native genome, with a tail-to-tail orientation relative to each other. All available data indicated that intact *gus* genes were present at both border regions.
- Fusion transcripts were detected by Northern blot analysis. 5' RACE showed that a majority, if not all (see Section 11.5.1), involved the left junction region, closest to the promoter of *PKT1*.
- The putative proteins encoded by *PKT1* and *PKT2* were found to be 3-ketoacyl-CoA thiolases. The two enzymes were essentially identical except for a 43 residues-long N-terminal extension in *PKT2* that includes a peroxisomal targeting signal. Comparative

analysis permitted the establishment of new, powerful signature sequences for members of this family of proteins.

- It was recognized that *PKT1* and *PKT2* are part of a family of genes in *Arabidopsis* that consists of at least four members.
- Northern blot analysis showed that transcripts hybridizing to *PKT1/PKT2* probes were present in all organs of *Arabidopsis* and at different developmental stages. That both genes were transcribed in the various organs and in embryos, and were independently regulated was demonstrated by RT-PCR.

***gus* expression and use of AtEN101 as a molecular marker**

- The patterns of *gus* expression were studied in plants growing under normal conditions and subjected to a variety of stimuli. GUS activity was found in all organs but under spatial and/or temporal regulation. In roots of seedlings, this activity was found to be influenced by several growth substances including IAA, 2,4-D and ABA. The effects of the growth substances were altered by the presence of sucrose in the growing medium.
- AtEN101 was also used as a molecular marker to determine the effects of *gnom* and *hydra* mutations on the patterns of *gus* expression in embryos. It was observed that homozygous cotyledonary stage embryos of either *gnom* or *hydra* displayed polarized and temporally regulated expression of the reporter gene, as was the case in *GNOM/AtEN101* and *HYDRA/AtEN101* embryos.

11.1.2 Isolation of a new morphogenetic mutant

- The library of *Arabidopsis* lines transgenic for pΔ*gus*Bin19 was screened for new mutants affecting seedling morphology. One such mutant was identified that had a characteristic dwarf phenotype, *bashful*.
- *bashful* embryos were often aberrant but these alterations did not have an endogenous cause. Where it was investigated, cells of *bashful* homozygous plants were found to be smaller than in wildtype, a characteristic that, among others was shared with several other dwarf mutants. The defect was rescued by growth in the presence of epibrassinolide.
- Segregation studies revealed that the *bashful* phenotype was linked to a kanamycin-resistance gene. However, although *gus* sequences are present in its genome, no GUS activity was observed in the mutant.

11.2 p Δ gusBin19 as a promoter trap

11.2.1 Cloning genes with promoter trapping

It is widely accepted that, at the cellular level, the developmental and adaptive changes of organisms involve as causation and/or consequence the differential expression of genes. It is also believed that the understanding of the underlying processes requires the comprehension of the nature, functions, regulation and interactions of the participating genes and their products (Section 1.1.2).

A number of methods and strategies have been devised through the years to identify and isolate genes involved in a variety of biological phenomena (Section 1.3). Among these, gene tagging is proving to be a powerful tool (Section 1.3.5). Among the tagging techniques promoter-trapping and related methods (enhancer- and gene-trapping) are particularly promising for their wide range of applications (Section 1.3.5.4): not only the tagging of genes with the trap immensely facilitates their isolation, but it permits the selection of genes of interest based on their apparent patterns of expression. This is a significant advantage when processes involving many genes, as is the case of embryogeny (Section 1.2.4), are being considered. In addition it allows for the isolation of genes whose transcripts are very low in abundance in a population of RNA molecules as a result of a very restricted mode of spatial and/or temporal expression. Furthermore, the selection of genes can be made independently of recognizable phenotypic alterations as a consequence of their mutation, although these may also occur.

In plants promoter trapping has been used to identify tissue-specific promoters, and a few genes have been reported cloned following this strategy. Teeri *et al.* (1986) used the promoterless *aph*(3')II gene as a reporter and selection marker next to the right border of T-DNA in the identification of two tissue-specific promoters. More recently, three tagged regions upstream from reporter genes were found to function as promoters active in nematode feeding structures (Barthels *et al.*, 1996). Genes cloned by promoter trapping include a nucleic acid helicase of the DEAH family in *Arabidopsis* (Wei *et al.*, 1997), and a D-ribulose-5-phosphate-3-epimerase induced at nematode feeding structures of *Arabidopsis* (Favery *et al.*, 1998). However, it is clear from the paucity of results that the full potential of promoter trapping for cloning of genes is yet to be transformed into a reality.

11.2.2 Validity of promoter traps as functional markers

Arguably the most important question with regard to promoter trapping concerns the reliability of reporter activity as an indicator of native gene expression. There are two aspects to this question, namely reporter activation as a result of genuine transcriptional/translational fusions, and the pattern of expression of a genuine fusion as an indicator of the expression of the tagged gene. Both aspects depend on the structure of the tag and the surrounding region in the mutated genome. Thus, reporter expression may result from the activity of a native promoter, and the tag

is expected to be inserted in the corresponding transcriptional unit (however, the reverse is not true). Alternatively, it may be driven by a cryptic promoter as was considered by Lindsey *et al.* (1993). One such promoter was found in a *N. tabacum* transformant of pPRF200, a T-DNA-based promoter trap, where the expression of the promoterless *gus* reporter gene was limited to developing seed coats (Fobert *et al.*, 1994). Cryptic promoters may not only be pre-existing in the tagged genome as they may also be newly formed with insertion of the tag. For example, in an *Arabidopsis* transformant of an *aph(3')*II gene-based promoter trap the activity of the reporter was demonstrated to be driven by a cryptic promoter formed by *cis*-elements from both a native promoter (acting in reverse orientation) and the T-DNA (Okresz *et al.*, 1998).

Promoter trapping in plants, usually involving T-DNA interposons, has generally been found to result in a significant proportion of transformants exhibiting reporter activity (Section 1.3.5.4). In *N. tabacum*, *N. plumbaginifolia* and *A. thaliana* transformants of a T-DNA vector with an *aph(3')*II reporter next to the right border about 30% of transcriptional fusions were detected (Koncz *et al.*, 1989). Kertbundit *et al.* (1991) observed fusion transcripts of *gus* in 54% of *Arabidopsis* transformants for the T-DNA promoter trap. Lindsey *et al.* (1993) observed similarly high frequencies of activation of *gus* in p Δ *gus*Bin19 transformants of three species. These were highest in different organ systems for each species ranging from 25% (12/48) in stems of potato, to 30% (28/94) in roots of *Arabidopsis* and up to 92% in flowers of 24 tobacco transformants. The high activation rates are surprising given the wide differences in genome size and repetitive DNA content of the species involved, and the average number of T-DNA loci (1-3) per transformant. This has prompted suggestions that T-DNA preferentially integrates into transcriptionally active or competent domains of the genomes (Herman *et al.*, 1990; Kertbundit *et al.*, 1991; Koncz *et al.*, 1989; Lindsey *et al.*, 1993). In contrast to these common findings, Fobert *et al.* (1991) found that in about 1000 tobacco transformants of the binary vector pPRF120 similar to that used by Kertbundit *et al.* (1991) only 5% expressed GUS reporter in leaves. For tobacco transformed with p Δ *gus*Bin19, 22% showed GUS activity in leaves (Lindsey *et al.*, 1993). More recently, only 7% of a new collection of 3,000 transgenic lines of *Arabidopsis* were found to exhibit reporter (GUS) activity (Favery *et al.*, 1998). The reasons for the substantial differences in activation of *gus* observed are not clear. It is possible that they may be related to variations in the transformation procedure and/or the structure of the T-DNA vectors used (see below). Regardless of these considerations, the very reduced proportion of repetitive DNA in *Arabidopsis* (15%; Meyerowitz, 1987; 1992) and its high gene density (one gene *per* 5 kb; Meinke *et al.*, 1998), indicate that interposon integration in this species has a high likelihood of occurring in or near transcription units.

Where the insertion of the promoter trap has created a genuine transcriptional fusion it may alter the regulation of the native promoter whose activity it reports. Changes in regulation can take place at all levels, including transcription, processing and transport of the transcript, translational and post-translational. In all cases the changes would result from structural

differences between the native and the tagged transcriptional units. At the level of transcription, modification of patterns of expression can arise from the functional disruption of *cis*-acting regulatory elements, and/or the addition of new ones contributed by the insertional mutagen. These problems have been recognized in a number of recent works (Taylor, 1997). For instances, reproduction of the correct pattern of expression of the *AGAMOUS* (*AG*) gene of *Arabidopsis* by promoter-*gus* gene fusion requires the inclusion a 3.8 kb intragenic region, located within an intron (Sieburth & Meyerowitz, 1997). This region was also demonstrated to be required for regulation of *AG* expression by other genes participating in flower development.

It is clear from the above that the structure of the integrated element is important in determining the activation of the reporter gene. Where they have been analyzed, the structures of T-DNA elements integrated in plant genomes vary substantially. T-DNAs often are truncated or rearranged, with multiple copies occurring as direct or inverted repeats at a single location. Jorgensen *et al.* (1987) found inverted repeats of T-DNA in 7 of 11 transformed tomato plants, where both the repeats were centered about a right border/right border or left border/left border junction. Grevelding *et al.* (1993) also found that a high proportion of T-DNAs may occur as multiple copies at a single site, up to 58% (14/24) being organized as right-border inverted repeat structures in *Arabidopsis* plants. The integration of multiple copies of a reporter gene may affect its expression, for example by undergoing transcriptional gene silencing mediated by methylation (Vaucheret *et al.* 1998). In a particular instance, several lines of tobacco transformed with pBI121 (where *gus* is under control of CaMV 35S promoter) were found to be distributed into two groups according to the level of GUS activity (Hobbs *et al.*, 1990). High expressers of reporter gene were all found to contain a single copy of T-DNA, whereas low GUS activity was only detected in those lines with multiple T-DNA copies. The level of activity was independent of the site of integration.

The integration pattern of T-DNA in a target genome has been shown to be influenced by a number of factors. Included is the strain of *Agrobacterium tumefaciens* used in the transformation procedure which influences the T-DNA copy number at the insertion site (Jorgensen *et al.*, 1987). The method of transformation also has a significant impact on the T-DNA copy number. In *Arabidopsis*, examination of 82 transgenic plants generated by different transformation experiments using several vectors showed that use of root tissue favoured the integration of single T-DNA copies (64%), whereas leaf-disk transformation resulted in 89% of transformants with multiple insertions, mostly as inverted repeats (Grevelding *et al.*, 1993). The method of transformation has also been suggested to have been a significant factor in the observed low level of reporter gene activation (7%) found by in the collection of *Arabidopsis* transformed *in planta* by Favery *et al.* (1998).

11.2.3 Promoter trapping in *Arabidopsis* with pΔ*gus*Bin19

It is clear from the above that many variables affect the patterns of integration of interposons and the activities of their reporter genes. It follows from this that each collection of transformants will have unique structural and activation characteristics which will be reflected in their degree of usefulness for identification and isolation of genes of interest.

As indicated earlier (Section 1.4.1) the library of pΔ*gus*Bin19-transformed lines (>2,000; Clarke *et al.*, 1992) is the largest T-DNA-based collection of promoter-trap *Arabidopsis* transgenics publicly available. It is therefore of interest to assess its value in promoter-trapping. The proportion of those pΔ*gus*Bin19-transformed lines expressing *gus* is quite high, although the frequency of those with a single T-DNA locus was about 50% (Lindsey *et al.*, 1993). In contrast, as indicated above, only 7% of the *Arabidopsis* promoter trap lines of Favery *et al.* (1998) were found to express *gus*. To date, molecular characterization of the pΔ*gus*Bin19 transgenic lines has been very limited. AtEN101 (da Rocha *et al.*, 1995; this work) and line atvt1 (Wei *et al.*, 1997) are the best characterized, and currently several other lines are being studied in Prof. K. Lindsey's group (S. Casson, Dr. M. Evans, K. Farrar, Dr. J.F. Topping, personal communications).

gus expression and LB region

The T-DNA in pΔ*gus*Bin19 includes a promoterless *gus* reporter in the left border repeat region followed by a *nos* terminator and, nearer the right border repeat, a *nptII* gene under the control of a *nos* promoter. The structure of this T-DNA at the left border places a difficulty and a possible solution to the activation of *gus*. The difficulty arises from the large distance (683 bp) separating the gene from the left border. This segment includes a total of 9 stop codons in frame with *gus*, the last of which is 270 bp upstream from the *gus* Met initiation codon, severely reducing the possibilities of formation of translational fusions (Section 4.2.6.3). Indeed, no large truncations at the left border have been found in all 8 cases analyzed (this work; Wei *et al.*, 1997; S. Casson, Dr. M. Evans, K. Farrar, Dr. J.F. Topping, personal communications). Rather, in all instances, only small deletions affecting the left border repeat of just a few bp up to less than 30 bp. It seems therefore that large segments separating *gus* from the native plant DNA are typical of pΔ*gus*Bin19 transformants. This contrasts with the commonly found deletion of left border sequences, to variable lengths which usually remove the entire repeat (*e.g.* Gheysen *et al.*, 1991).

On first analysis, the large distance to the left border and the stop codons found in the sequence upstream from *gus* have a strong potential to reduce or eliminate the expression of the reporter, even when transcriptional fusions are created. This is particularly true where insertion of a nearly intact T-DNA occurs in a protein-encoding exon, and premature translation termination should ensue. However, all 8 structurally characterized left border/plant DNA junctions analyzed derive from 6 lines expressing *gus*. This apparent paradox can be partially

explained by the splicing out of large sections of the 683 bp LB segment. The possibility of splicing was recognized by the identification of three potential 3' splice sites (at positions -101, -276 and -700) in the region (Section 4.2.6.3). Use of these sites would be expected to facilitate the expression of the downstream *gus*. The finding of mature fusion transcripts resulting from the splicing of the first exon of *PKT1* to -101 at the LB in AtEN101 demonstrated that the predicted 3' splice site at this position is functional (Section 6.2.4). Furthermore, this finding shows that functional transcriptional fusions can be created even in those instances where the T-DNA has inserted in an intron, as is the case in AtEN101. In line atvt1, where GUS activity is observed, the T-DNA is also believed to be present in an intron of the tagged gene (Wei *et al.*, 1997). In a third line, the T-DNA is apparently inserted in an exon (S. Casson, personal communication). The insertion sites for the other lines being studied are unclear although in two instances they are unlikely to occur in protein coding regions (Dr. M. Evans, personal communication).

Because for RT-PCR of fusion transcripts in AtEN101 amplification was performed with primer LB-267 (Section 6.2.3), splicing to positions -276 and -700 could not have been found. Hence their possible role in processing of fusion transcripts remains to be established. It is noteworthy that splicing to -700 from a 5' splice site located in the native genome would eliminate all of the left border region and potentially create protein fusions to *gus* 14 bp downstream from its AUG codon. Analysis of the 623 bp sequence with NetPlantGene V2.1 (Hebsgaard *et al.*, 1996) also identified the 3' splice site at -101 but not those at -276 and -700, or any 5' splice sites (not shown). In AtEN101, such splicing is unlikely to take place at least on a majority of *gus* (fusion) transcripts: their length was estimated to be 2.5-2.8 kb (Section 8.3) and should include the *gus-nos* region (~1.9 kb) and 5' RACE of fusion transcripts indicated that (most) had the same 5' end of *PKT1* transcripts (see below 11.5.1).

pΔ*gus*Bin19 and high frequency of GUS expressing transformants

An interesting question is raised concerning translation of fusion transcripts where GUS activity is detected if no splicing to -700 takes place, particularly where insertion occurred within a protein coding region. There may be situations where potential polycistronic transcripts are generated. Some instances of eukaryotic polycistronic messages are known, including a suspected case of a plant nuclear transcript (Blumenthal, 1998). Koncz *et al.* (1989) also observed translation of dicistronic transcripts in transformants of tobacco including that of a promoter trap vector. Therefore, the possibility that a functional internal ribosomal entry site exists in the region of the first methionine codon of *gus* has to be considered and would be of great interest to explore.

As was shown in Section 6.3, at both junctions with the native genome the T-DNA in AtEN101 consists of left border regions, each apparently including an intact *gus* gene. This was the first demonstration of a dyad-type organization of T-DNA copies in pΔ*gus*Bin19

transformants (Ferreira da Rocha *et al.*, 1996). In one of two other lines for which information is available on the T-DNA organization both junctions with the native genome also consist of left border regions (Dr. M. Evans, personal communication). A single T-DNA copy is found in the third line. The occurrence of two left border regions at the T-DNA/plant DNA junctions in a given insert ensures that reporter gene activity is independent of the relative orientation of the T-DNA. Although the data available is not statistically significant it does show that these arrangements occur non-rarely in transformants of p Δ *gus*Bin19, and so may contribute to the high frequency of *gus* activation in this species.

Another factor potentially contributing to the high frequency of GUS expression in transformants is the activity of cryptic promoters. Relevant in this respect was the finding that the E1.4E1.5 fragment, thought to include promoter elements of *PKT2* (see below 11.3.2), was capable of driving the expression of *gus* in pPRP19-4 regardless of orientation.

11.2.4 The use of *gus* as a reporter gene

The characteristics and the range of substrates available for GUS, making its assay simple and highly sensitive, and its detection *in vivo* possible (Jefferson *et al.*, 1987), have transformed it into a preferred reporter enzyme. The sensitivity of GUS assays is partially due to the very high stability of the enzyme which has been estimated to have an half-life of ~50 hours in many plant cells (Jefferson *et al.*, 1987). While this is advantageous for detection of activity where its level is very low, it is highly undesirable where a reduction or cessation of transcription takes place. In these instances a substantial lag period (several days) will mediate between the change in transcription and an eventual change in gene activity.

This characteristic of GUS has a significant negative impact on its usefulness as a reporter. For example, it is impossible to study changes that may occur on a daily/circadian clock-linked cycle. Study of the effects of external stimuli on GUS activity in AtEN101 also illustrates this difficulty: the stability of the enzyme is likely to have delayed substantially the observed repression of activity in seedling roots following the addition of ABA and auxins (Section 9.3.2).

11.2.5 Final considerations

In conclusion, the study of AtEN101 showed that in p Δ *gus*Bin19-transformed *Arabidopsis* lines reporter activity may result from expression of fusion transcription units. Processing of the primary transcripts can involve splicing out of portions of the left border region eliminating structural obstacles to translation. This is likely to contribute to the high frequency of reporter activity in the transgenic lines. Reporter activity may also derive from cryptic promoters present in the *Arabidopsis* genome or newly created by insertion of the T-DNA.

The demonstrated occurrence of inverted repeats of T-DNA copies is likely to contribute to the elevated number of transformants exhibiting GUS activity. A practical

consequence of the presence of inverted repeats, is an increased difficulty in establishing a relationship between *gus* activity and the expression of a putative tagged gene. An instructive example was the finding that the reporter activity of a (different) T-DNA-based promoter trap, inserted into an intron of an *Arabidopsis* growth arrest response gene, was driven by a cryptic promoter and was not related to that of the tagged gene (Okresz *et al.*, 1998). From the limited data available, the p Δ *gus*Bin19 T-DNA is frequently found in non-coding regions, at least for GUS-expressers. Given the general intactness observed at the LB, if splicing to -700 does not occur, it may be that most of the GUS-expressers will have the reporter gene inserted at a considerable distance from protein encoding regions. These observations highlight the need to characterize other GUS-expressing lines in the same collection of transformants.

In spite of its limitations, promoter trapping is a very useful tool in the isolation of genes of interest. This is illustrated by a T-DNA-linked mutation affecting plant shape, *bashful*, which demonstrates the value of promoter-traps in the isolation of genes via their mutagenic potential. In this instance, although no *gus* expression was observed, the results indicate that the affected function is involved in hormonal control of cell shape (see below, Section 11.6). Complementary, it allows for the isolation of genes based on a simple assay of reporter expression and independently of the existence of an associated phenotype. This was illustrated here by the isolation of *PKT1* and *PKT2* (which apparently are apparently not functional in AtEN101) in the absence of a phenotype. The isolation of these two genes demonstrates the substantial value of promoter trapping for the identification of genes which would likely go undetected in conventional genetic screens.

11.3 Two overlapping *PKT* genes in *A. thaliana*

11.3.1 Unusual genic organization of *PKT1* and *PKT2*

The isolation of cDNAs for the two distinct but nearly-identical transcripts and comparative analysis with the genomic region tagged in AtEN101 demonstrated that their transcription units overlapped for most of their length (Chapter 5). That the two transcripts derive from distinct genes rather than differential splicing is indicated by a strong body of evidence. A scenario involving differential splicing would require their sharing of the 5' termini. However, 5' RACE only added 105 nt to the known 5' limit of the *PKT2* transcription unit (in pTB-6) placing it 623 bp downstream from that determined for *PKT1* also by 5' RACE (Section 5.4). The total known length of exon sequences determined for *PKT2* is then 1691 bp. This length is similar to that estimated from Northern blot analysis for the *PKT1*/*PKT2* transcripts, 1.6 kb (Section 9.2.1), and transcripts of plant peroxisomal thiolase genes in *M. indica*, *Cucurbita* sp., and *C. sativus* (1.7 kb; Preisig-Müller & Kindl, 1993; Bojorquez & Gomez-Lim, 1995; Kato *et al.*, 1996). Also, no potential splice sites were found in the 623 bp segment separating the first exons of the

two transcription units, and the 0.5 kb E1.5E1.4 segment overlapping with it, directed transcription of *gus* in transient assays with plasmid pPRP19-3. RT-PCR with prTHIO-3F and primers from the common sequence failed to reveal the presence of *PKT2*-type transcripts where exon 1 of *PKT1* was present. However, final proof of the existence of two distinct genes will require mapping of the 5' end of the *PKT2* gene.

The organization of *PKT1* and *PKT2* in *Arabidopsis* constitutes a rare example of gene overlap in plants. In a number of species the genes encoding pyruvate orthophosphate dikinase (*pdk*) have a similar arrangement (*e.g.* Sheen *et al.*, 1991; Rosche & Westhoff, 1995). The gene encoding the chloroplastic form of the enzyme shares with that for the cytosolic form the same exon/intron structure except for an additional first exon encoding the chloroplast targeting signal. The use of alternative promoters that result in different lengths of the first exon is more common, as exemplified by genes for the *Arabidopsis* 3-hydroxy-3-methylglutaryl coenzyme A reductase (Lumbreras, 1995).

Given the higher conservation of *PKT2* relative to other thiolases and their genes (Chapters 7, 8) it is likely that *PKT2* predates evolutionarily *PKT1*. The latter possibly resulted from mobilization of sequences with promoter and transcript encoding functions found upstream from *PKT2*. Its selection might have resulted from the enabling of expression of a *PKT1*/*PKT2*-type thiolase under a greater variety of regulatory controls (resulting from the use of two promoters).

11.3.2 The promoter regions of *PKT1* and *PKT2*

PKT2

The size of the *PKT1*/*PKT2* transcripts and related transcripts of other plant species (see above) indicate that the transcription initiation site (TIS) of *PKT2* is near (if not coincident with) the known limit of the transcription unit defined by 5' RACE. A number of matches to consensus sequences for *cis*-regulatory elements of transcription were found in the region, most with a low to very low probability of random occurrence (Section 8.4.2). Included is a single TATA-box at -275. Use of this TATA-box would be expected to originate a processed transcript of over 1.9 kb, but transcripts of this size were not detected by Northern blot hybridization. Also, although the TATA-box is located only 62 bp downstream from E1.5, the E1.5E1.4 segment was sufficient to direct expression of *gus* in transient assays. Together these results indicate that this match to the TATA-box consensus specifies the TIS of *PKT2*. Rather, it seems that the promoter is TATA-less.

It is interesting to note the clustering of several matches to *cis*-acting elements in several groups. One cluster is located upstream from the TATA-box (9-24 bp distances; Hexamer ath4 -562, ELRE core -546, I-boxes -531 and -502, GT-1 box -494), another overlaps

it (1-26 bp; Myb-st1 -355, GT-1 boxes -333 and -310, Myb-st1 -304/-301, GT-1 box -270), and a third is further downstream (17 bp; G-box -166, LTRE/amy3 -144). In the *Arabidopsis RBC S-1A* promoter several of these elements (a GT-like box, two I-boxes and a G-box), required for activity, are also separated by short segments (9-20 bp; Donald & Cashmore, 1990). Clustering allows for short-range interactions of DNA-binding activities that may recognize these elements. The sequence CCCACGTGGCCCGG at -164 deserves particular consideration as it includes a G-box core and several very low probability (10^{-4} to 10^{-5}) matches to elements (ABRE) that mediate regulation by ABA in numerous gene promoters in plants, including *Arabidopsis*, as well as to other elements (Section 8.4.2). Whether any of the matches to *cis* elements function in the *PKT2* promoter remains to be determined. However, the observed clustering and their low probability of occurrence suggest that they may participate in the regulation of *PKT2* expression. If functional, the possible roles of the various elements cannot be inferred from their known functions in other genes. Nevertheless, it is tempting to speculate that the potential ABRE element may have a role in repressing transcription of the gene during seed maturation/desiccation stages regulated by ABA, when storage lipids are being accumulated.

PKT1

The 5' limit of mP18 was defined by two clones obtained by 5' RACE of *PKT1* transcripts (Section 5.4.4) and coincided with that of the longest fusion transcript extending into the region (Section 6.2.4). This indicates that either this position is the TIS for *PKT1* or is a preferred target for degradation. A TATA box is found closely upstream (-28 bp), within the normal range of distance to TIS (Breathnach & Chambon, 1981), and it is therefore possible that the site is or maps near to TIS. The estimated size of the resulting transcripts would match that of 16S chloroplast rRNA, as observed in Northern blot analysis (Section 9.2.1).

Although two other TATA boxes were found at -375 and -770 no potential splice sites, hence exons, were found in the region upstream from mP18 (Section 8.2.3). More importantly, the transient expression assays with pPRP18.28-106 showed that the sequence from E5.1 (+26) to -360 was sufficient to drive transcription of the reporter *gus*. However, a role of the small segment of TnΔXR (168 bp) present in the construct cannot be excluded.

Noteworthy is also the distribution of most matches to *cis*-acting elements which are concentrated in the segment up to -450. The probability of chance occurrence of many of these elements was relatively high ($P > 0.1$), however their clustering within the 1.2 kb region may be significant (see above, "*PKT2*"). A cluster formed by some elements with low probability is found at about 0.1-0.2 kb upstream from mP18 (H-boxes -171 and -136, and the I-box -158). In addition the I-box in this cluster is contained within a L-box ($P = 2.528 \times 10^{-4}$), first described in phenylalanine ammonia-lyase genes (Logemann *et al.*, 1995). A close match to a L-box is also

found 69 bp downstream (CCTCATCTAAC). Together the data are in agreement with and reinforce the notion that TIS is located at or near the 5' end of mP28.

The possible roles, if any, of the mentioned matches to *cis*-acting elements are unknown. However, given the predicted enzymic nature of PKT2, the occurrence at -280 of a match to the Sucrose Responsive Element (SURE; AATAGAAAA; $P_{1-350}=6.76 \times 10^{-3}$) may be relevant. The SURE element was first functionally described in the potato patatin promoter where it is involved in the sucrose-triggered induction of transcription, and is conserved in other sucrose-responsive genes (Grierson *et al.*, 1994). Also found 78 bp upstream is a sequence (TATATATATATATGTATAGA) with strong similarity to box-A (TATATATATATTATAT-AATA) of the same patatin promoter which has a repressing role on expression of the gene in specific tissues of potato (Grierson *et al.*, 1994). Curiously, the effects of ABA and the tested auxins on GUS activity in AtEN101 were modulated by the presence of sucrose in the medium. It would be interesting to investigate if the SURE and box-A elements in the promoter region of PKT2 are involved in mediating the observed sucrose effects.

11.4 PKT1 and PKT2 as peroxisomal thiolases

11.4.1 Are PKT1 and PKT2 real proteins?

Overwhelming structural comparative evidence presented in Chapter 7 shows that the putative proteins PKT1 and PKT2 belong to the class of peroxisomal 3-ketoacyl-CoA thiolases. The high primary structure similarity to known thiolases, and secondary structure and 3D modelling indicate the nature of the enzymic function of the proteins: 3-ketoacyl-CoA thiolases catalyze the last step of the β -oxidation pathway of fatty acids, the cleavage of $C_{\alpha}-C_{\beta}$ of an acyl_{*n*}-CoA in a thiolysis reaction with CoA to form acetyl-CoA and an acyl_{*n-2*}-CoA.

Although no direct evidence was gathered for the existence of PKT1 and PKT2, the available information points for the likely translation of their transcripts. First, the sequence similarity of PKT1 and PKT2 to other thiolases is uninterrupted throughout their length indicating that they are not pseudogenes (please see below for discussion on the N-truncation in PKT1). Secondly, primary sequence analysis showed that the heptanucleotides containing the first methionine codons for PKT1 and PKT2 in mP18 and mP19 were among the most frequently found at translation initiation sites in *Arabidopsis* genes (Section 7.2). Thirdly, search for possible strong elements of secondary structure in the vicinity of the putative translation initiation region of the transcripts failed to reveal any that might mask the initiation sites. Together, the data indicate that translation can be initiated in mP18 and mP19 transcripts. It further testifies to the existence of evolutionary pressure to preserve the genes, likely resulting from their metabolic role. The challenge remains to demonstrate the existence of both forms of

the protein. In this respect, the *c-myc* tagged variants of the corresponding cDNAs (Section 7.8.3) may prove useful.

11.4.2 Intracellular targeting of PKT1

The organellar import of most peroxisomal matrix proteins, like thiolase, is dependent on one of two well-characterized targeting signals (Gietl, 1996): a 3 residues-long C-terminal PTS1 signal, the most commonly used, or a longer N-terminal PTS2 signal that is present in all peroxisomal thiolases. Therefore the lack of a peroxisomal targeting signal (PTS1 or PTS2) in PKT1 (Section 7.8.2) is surprising and unique, raising the possibility that this protein might have a distinct intracellular site of action. In a few instances other ill-defined PTS signals have been identified in proteins other than thiolases (Gietl, 1996), and, even if unlikely, it cannot be ruled out that one such signal may be present in PKT2.

Two systems of β -oxidation, a mitochondrial one and another peroxisomal, are present in most groups of organisms. In plants no elements of a mitochondrial system have ever been found although there is controversy about its possible existence, *e.g.* (Dieuaide-Noubhani *et al.*, 1992). It could be tempting to suggest that PKT1 could be one such element, particularly in view of the structure at its N-terminus which has properties in common with some mitochondrial targeting signals (Section 7.8.2). But this is very unlikely, not only because of lack of other evidence for mitochondrial components, but principally in view of the recognition that in organisms where both systems exist, the mitochondrial and peroxisomal forms are structurally quite distinct from each other (*e.g.* Arakawa, 1987; Igual *et al.*, 1992).

Rather it is suggested that PKT1 is imported by peroxisomes, in spite of the absence of a targeting signal, by a piggy-back mechanism. A truncated form of yeast thiolase lacking the 16 N-terminal residues is co-imported into peroxisomes together with a full-length thiolase following dimerization of the protein (Glover *et al.*, 1994). Dimerization involves the activity of Pex20p proteins and is required for the import in a PTS2-independent manner (Titorenko *et al.*, 1998). Therefore, PKT1 can be imported into the peroxisomes via heterodimeric association with other, PTS2-bearing thiolase units.

Given that the modelled association of PKT1/PKT1 does not involve interactions between PTS2 elements (Section 7.7), as was demonstrated for the truncated and normal forms of the yeast thiolase (Glover *et al.*, 1994), PKT2 can heterodimerize with PKT1. It was possible to construct a model of an heterodimer of PKT2 and PKT3 by homology-threading, indicating that their association can take place (data not shown; very similar to that in Fig. 7.9C). PKT3 is expected to be present in high excess relative to the other peroxisomal thiolase forms (see Section 11.4.4), and therefore is likely to ensure an efficient peroxisomal import of PKT1.

11.4.3 The peroxisomal targeting signal PTS2

Very recently a new (non-absolute) consensus was derived for PTS2 consisting of the nonapeptide: -R/K-X₆-H/Q-A/L/F, where X₃ is normally an aliphatic residue (Flynn *et al.*, 1998). During the course of the work presented here an effort was also made to characterize PTS2 by comparative analysis. On the basis of a larger sample of sequences from PTS2-bearing proteins (33 including all those considered by Flynn *et al.* (1998) the following decapeptide consensus for PKT2 was arrived at: -Polar-R-X₃-Hydrophobic-X₂-H/K/Q-A/L/F. Only two sequences (human and guinea pig alkyl-DHAP synthases) were found not to follow this convention in that the first residue in their decapeptides is Alanine. Flynn *et al.* (1998) suggested that the functional PTS2 is not solely defined by the consensus residues but also by the structural motif where it is inserted. Such dependency may explain the deviations from the consensus observed for the two alkyl-DHAP synthases.

Another conserved structural feature among PTS2-targeted peroxisomal enzymes is the presence of a proline residue in the C-vicinity of the signal. The comparative analysis also revealed that only four enzymes have prolines separated from the signal by more than 7 residues (Fig 11.1). The proline residue may have a role in ensuring the availability of the PTS2 residues for interaction with the peroxisomal import machinery. This hypothetical role of proline may be achieved by its preventing the extension of downstream elements of secondary structure into PTS2. If such extensions were to occur they might interfere sterically with PTS2. Also, they might involve the physico-chemical potential of the PTS2 residues in intramolecular interactions that would impede their participation in the targeting signal. Proline is ideal to fulfill this role as its unique structural characteristics cause it to generally occur in loop regions or, more seldom, at the ends of secondary structure elements, usually α -helices, where it acts as a "breaker" (Richardson & Richardson, 1989). Even when present in more internal positions in α -helices, proline induces a bend in the structure. Mathieu *et al.* established that in the crystallized dimer of the *S. cerevisiae* thiolase the first 24 residues are disorganized (Mathieu *et al.*, 1994; 1997). Examination of the tertiary structure of this enzyme shows that the first proline following PTS2 is involved in a B-turn, two residues away from the first β -strand.

11.4.4 A family of PKT genes

In addition to the two PKT genes characterized here, two additional genes encoding peroxisomal thiolases are known (Section 10.3), including PKT3 (*PED1*, Hayashi *et al.*, 1998) and PKT4 (for which two corresponding ESTs have just been reported). All genes share a similar exon/intron organization, and both their protein-encoding portions, and transation products are very similar ($\geq 70\%$; Section 10.4.2). The higher degree of similarity between the pair PKT3 and PKT4 is indicative of a more recent duplication than that originating the ancestor of PKT1/PKT2 and of PKT3/PKT4.

			*	20	*	40	
1) A.t.-PKT2	:	-----	MERAME	QKILLRHLP	-----	:	17
2) A.t.-PKT4	:	-----	MEKATER	QRIILLRHLP	-----	:	17
3) A.t.-PKT3	:	-----	MEKATER	QRVILLEHLP	-----	:	17
4) B.n.-PKT	:	-----	MEKAME	QRVILLEHLP	-----	:	17
5) M.i.-PKT	:	-----	MEKAIN	QQVILQHLP	-----	:	17
6) C.s.-PKT	:	-----	MEKAIN	QSIILLHHLP	-----	:	17
7) C.m.-PKT	:	-----	MEKAIN	QSIILLHHLP	-----	:	17
8) C.t.-PKT	:	-----	MDRLNQL	SGQLKE	-----	:	13
9) Y.l.-PKT	:	-----	MDRLNNL	ATOLEQNP	-----	:	15
10) S.c.-PKT	:	-----	MSQRLQSI	KDHLVESAMGKGESKRKNSLLEKRP	-----	:	33
11) R.n.-PKT-A	:	-----	MSESVGR	TSAHRLQVVLGHLAGRP	-----	:	25
12) R.n.-PKT-B	:	-----	MHRLQV	VLGHLAGRSESSSALQAAE	-----	:	25
13) H.s.-PKT	:	-----	MQR	LQVVLGHLIRGP	-----	:	14
14) C.s.-gMDH	:	-----	MQPIPD	VNRIARISAHLP	-----	:	20
15) O.s.-gMDH	:	-----	MEDAAAA	ARMERIASHLP	-----	:	20
16) M.s.-gMDH	:	-----	MEPN	SYANSRITRIASHLP	-----	:	20
17) G.m.-gMDH	:	-----	MEANS	GASDRISRIAGHLP	-----	:	20
18) B.n.-gMDH	:	-----	MPHK	RIAMISAHLP	-----	:	15
19) C.l.-gMDH	:	-----	MQPIPD	VNRIARISAHLP	-----	:	20
20) C.p.-gMDH	:	-----	MKPIPD	VNERIARISAHLP	-----	:	20
21) H.p.-AMO	:	-----	MERLRQ	IASQATAASAAE	-----	:	18
22) T.b.-ALD	:	-----	MSK	VEVILTQLP	-----	:	13
23) H.s.-PAHX	:	-----	MEQLRA	AARLQIVLGHLGTP	-----	:	20
24) M.m.-PAHX	:	-----	MNLTR	AGARLQVVLGHLGRP	-----	:	20
25) H.s.-ADS	:	-----	DPDP	DRAGRLRVLSGHLGRP	-----	:	22
26) G.p.-ADS	:	-----	DPDQ	DRAGRLRVLSGHLGRP	-----	:	22
27) R.n.-PEX	:	-----	VANQS	QGRDLFRATQHACMLLRYLLES	-----	:	28
28) H.p.-PEX	:	-----	MQPWY	HKLGRQGRQLEAQWQTDAAE	-----	:	25
29) A.t.-AAT	:	-----	FSSSS	SDRIGAILLRHINSRSDSDNLSSLYASE	-----	:	34
30) G.m.-AAT	:	-----	SSSSP	PCDRRLNTLARHFLP	-----	:	20
31) Ph.-OXI	:	-----	EHDTQ	QALRRIQKLSLHLLQP	-----	:	21
32) C.s.-OXI	:	-----	EDESQ	AAARRIERLSLHLP	-----	:	20
33) C.m.-HSP	:	-----	SMRKS	NHVSSRTVFFGQKLGNSAFAE	-----	:	26
34) C.s.-HSP	:	-----	SMRKS	NNVSSRVSFFGQKLGNSAFAE	-----	:	26
35) C.s.-CS	:	-----	MPTDM	ELSPSNVARHRLAVLAHLAASLEP	-----	:	31

Figure 11.1 Alignment of PTS2 sequences from peroxisomal proteins

The matches to the proposed consensus (see text) are highlighted with grey or black (total conservation) background. Also shown in black background is the PTS2-proximal Pro residue. Sequences lacking a Met at the N-terminus, are shown truncated. Codes for species as in Figure 7.5. Additional species include: C.s.- *Citrullus lanatus* (watermelon); G.p.- guinea pig; M.s.- *Medicago sativus* (alfalfa); Ph-Phalaenopsis. Additional codes for proteins include: gMDH-glyoxysomal malate dehydrogenase; AMO-amine oxidase; ALD-aldolase; ADS-alkyl-dihydroxyacetonephosphate synthase; AAT-aspartate amino transferase; OXI-acyl-CoA oxidase; HSP-heat shock protein 70; CS-citrate synthase. References for sequences 1-13 can be found in the legend to Figure 7.5. Other references: 14) Kim & Smith (1994); 15) Kaminaka & Tanaka (1996); 16) Miller *et al.* (1997); 17) Guex *et al.*, (1995); 18) Witt *et al.* (1995); 19) Gietl (1990); 20) Kato *et al.* (1998); 21) Bruinenberg *et al.* (1989); 22) Clayton (1985); 23 and 24) Mihalik *et al.* (1997); 25) de Vet *et al.* (1997b); 26) de Vet *et al.* (1997a); 27) Passreiter *et al.* (1998); 28) Waterham *et al.* (1994); 29) Schultz & Coruzzi (1995); 30) Gebhardt *et al.* (1998); 31) Do & Huang (1997); 32) Hayashi *et al.*; 33) Wimmer *et al.* (1994); 34) Preisig-Mueller *et al.* (1993); 35) Kato *et al.* (1996b).

The existence of several genes encoding isozymes of thiolase testifies to the metabolic importance of the enzyme. The occurrence of the three forms of the isozyme (PKT1/PKT2, PKT3 and PKT4) is likely to translate an adaption to the conversion of a variety of substrates. The control of expression of *PKT1/PKT2* by two distinct promoters potentiates its utilization under a greater variety of conditions.

ESTs for a fifth gene with similarity to the *PKT* genes were also identified (Section 10.3). Although the partial sequence of the protein encoded by this gene was significantly more distinct of all C-terminal sequences (see Fig. 10.6), it contains the three thiolase signatures mapping to this region of the enzymes. This also includes one of the enzymatically important Cys residues. It seems therefore that the fifth gene also encodes a thiolase, but evolutionarily more distant.

11.4.5 The specific functions of PKT1 and PKT2

Whereas the general enzymic activity of PKT1 and PKT2 as thiolases is well established, the existence of a family of at least four genes encoding these enzymes in *Arabidopsis* indicates that each should have a specific or preferred role. The knowledge of the expression patterns of the various genes may be helpful in determining the roles of the various isozymes and their genes.

Search of the EST database had identified 21 clones derived from *PKT3* transcripts but only 2 corresponding to *PKT1/PKT2* transcripts (Section 10.3.2). Because the EST clones were prepared from a mixture of organs of plants grown under various conditions it can be said that generally expression of *PKT3* is considerably higher than *PKT1/PKT2* (and *PKT4*). Indeed, the difference has been estimated to reach 2 orders of magnitude (Dr. M. Hooks, personal communication). Although proportionally a minority, *PKT1* and *PKT2* transcripts are found in all organs and at different developmental stages of *Arabidopsis* (Section 9.2). The Northern blot hybridization results seem to corroborate this. However, these latter results should be viewed with caution given the presence of two different size transcripts only in (some) young germinating seedlings (Section 9.2.1), but not elsewhere. Because of the high similarity of *PKT1/PKT2* to *PKT3* transcripts and difference in levels of expression it cannot be ruled out that one of the hybridizing transcripts corresponds to *PKT3* (performing parallel experiments with probes for each transcript, would greatly help in clarifying the expression patterns of all four known *PKT* genes).

Regardless of the possible involvement of *PKT1/PKT2* during mobilization of lipid reserves during germination (suggested by the apparent induction seen in young seedlings; Fig. 9.3), it is clear that both participate in β -oxidation throughout the plant and at various stages of its life cycle. β -oxidation is not limited to glyoxysomes of seedlings but also takes place in leaf and non-specialized peroxisomes of mature plants, occurring in a variety of organs including spinach leaves, mung bean hypocotyls, potato tubers and in roots (Gerhardt, 1983). Its metabolic importance is multiple: in addition to supplying acetyl-CoA for gluconeogenesis (*e.g.*

in seedlings) and for the anaplerotic carbon flux of citric acid cycle intermediates (Tolbert, 1981), it would be expected to participate in the normal metabolism (turnover) of structural lipids of all cells. In addition, in senescent pumpkin cotyledons, leaves and petals, and in starved (detached) leaves and roots β -oxidation induction is accompanied by that of malate synthase gene transcription (Graham *et al.*, 1992). Isocitrate lyase gene transcription was also induced in the senescent organs (unpublished results cited in Graham *et al.*, 1992). This enzymic complement allows for gluconeogenesis to take place. However, sugar starvation leading to induction of β -oxidation may not result in gluconeogenesis: Dieuaide-Noubhani *et al.* (1997) found that sugar-deprivation of excised maize root tips results in lipid (and protein) mobilization accompanied by increase in malate synthase activity but not isocitrate lyase. In this instance the lipids are therefore thought to be mobilized for respiration. In summary there is a growing body of evidence indicating the existence of multiple roles for the pathway, and that the various mechanisms involved in lipid mobilization and the final destination of the β -oxidation product are under metabolic control (*e.g.* Graham *et al.*, 1992).

The β -oxidation pathway is also used in the regulation of the cellular lipid composition. For example, Hooks *et al.* (1999) showed that over-expression of a medium-chain acyl-acyl carrier protein (ACP) thioesterase MCTE in *Arabidopsis*, by placing it under the control of the CaMV 35S promoter, does not result in accumulation of laurate in leaves. Therefore, the pathway appears to have a "homeostatic" role in this respect. Curiously, no increase in transcription or enzymic activity of β -oxidation enzymes was observed. This was taken as evidence that the endogenous β -oxidation was sufficient to maintain the normal equilibrium in lipid composition (Hooks *et al.* 1999). In similar transformants of *B. napus* where thioesterase activity is high in leaves no accumulation was observed (Eccleston *et al.*, 1996). This was suggested to be due to an increase in β -oxidation. Also, although in developing embryos of these transformants there is accumulation of laurate (C12:0) in triacylglycerols, there is also an induction of β -oxidation (Eccleston *et al.*, 1998).

The existence of multiple roles for β -oxidation may lie at the base of the existence of several genes encoding thiolase by allowing independent regulation of each to respond to various, distinct demands. This may explain the maintenance of the two genes, *PKT1* and *PKT2*, encoding almost identical proteins. Another, and non-mutually exclusive possibility is that the different isozymes have distinct substrates. The nature of these substrates is quite varied, something not always suitably recognized, and includes fatty acids of widely varying lengths and of more or less complex structure (*e.g.* branched-chain fatty acids), dicarboxylic acids, prostaglandins (in animals), and xenobiotics with acyl side chains. Although β -oxidation enzymes have a broad range of substrates, they tend to have preferences (Tolbert, 1981). It may be that some compounds can only be metabolized by one or other of the isozymes. Significantly, Hayashi *et al.* (1998) were able to select *ped1*, an *Arabidopsis* mutant defective for the thiolase gene here referred to as *PKT3*, by growth of mutant lines in toxic levels of 2,4-

DB. This compound is transformed by β -oxidation into the toxic 2,4-D and mutants (such *ped1*) resistant to 2,4-DB but not 2,4-D are expected to have defects in the pathway (Hayashi *et al.*, 1998). Assuming that the Landsberg ecotype on which *ped1* was isolated has the same complement of thiolase genes as C24 and Columbia, the isolation of the mutant demonstrates that the other thiolase isozymes were incapable of metabolizing 2,4-DB. Whether this resulted from a low level of expression or incapacity of their protein products to use 2,4-DB as a substrate is not known.

Regardless of the particular functions of PKT1 and PKT2, the observed modulating role of sucrose on the effects of ABA in the expression of *gus* shows that the expression of their genes is sensitive and responsive to changing physiological conditions. Because sucrose is the final product of gluconeogenesis, as well as a starting point for glycolysis and formation of acetyl-CoA, it is possible that this modulation takes place via a feedback mechanism (see below Section 11.5.1).

11.4.6 New signature sequences for thiolases

The conservation of three blocks of sequence in thiolases was recognized (Section 7.5) that were developed into new signatures for this family of enzymes (Section 7.9). The existence of the blocks is indicative of their importance for the function of the enzymes. For this reason the localization of the blocks in the yeast peroxisomal thiolase 3D structure and the model built for PKT2 was examined. All three were found to occur in the core of the proteins.

Thiolase subunits are composed of three domains: an N-terminal core domain is linked to the C-terminal core domain by a long loop domain (Mathieu *et al.*, 1994; 1997). Each domain is about 120-140 residues long. The two core domains share the same topology consisting of a mixed β -sheet of five strands sandwiched between helices, an arrangement described as $\beta 1\alpha 1\beta 2\alpha 2\beta 3\alpha 3\beta 4\beta 5$ (see Figs. 7.9 and 7.10). N $\beta 4$ and N $\beta 5$ (where N refers to N-domain, using the terminology of Mathieu *et al.*, 1994) are separated by the loop domain. Together the β -sheets of both core domains frame the shallow pocket containing the active site. Block A encompasses N $\beta 4$, which is the central β -strand of the N-terminal domain. N $\beta 4$ is parallel to N $\beta 2$ and antiparallel to N $\beta 5$, with which delimits the loop domain. N $\beta 4$ therefore seems to have a structuring role in the β -sheet region and in placing/orienting the loop domain. This is likely to explain the conservation of this block of residues. Similar considerations can be made with regard to block B which corresponds to most of N $\beta 5$. This is the central strand in the antiparallel β -sheet region formed together with N $\beta 1$ and N $\beta 4$.

Block C of conserved residues is located in the C-terminal core domain. It extends from the central region of the loop linking C $\alpha 1$ to C $\beta 2$ to the N-terminal residue of C $\alpha 2$. It includes the short (4 residues-long) C $\beta 2$ and the short loop separating it from C $\alpha 2$. C $\beta 2$ is the central strand in the parallel β -sheet region of the C-terminal core region which it forms with C $\beta 3$ and C $\beta 4$. It is likely that block C residues are important in forming the parallel β -sheet region. In

the yeast peroxisomal enzyme, Gln₃₄₁ in C β 2 is known to form a salt bridge with Arg₃₈₃ in C α 3. Equivalent residues are well conserved in thiolases (Mathieu *et al.*, 1994; this work). This interaction is conceivably relevant for positioning of C α 3.

Conservation of block C may also be linked to the close proximity of the four residues in the loop C β 2 to C α 2 to the active site region ($< 9 \text{ \AA}$). Indeed, two of these residues (Arg₃₆₂ and Ala₃₆₄ in PKT2) are less than 5 \AA away. This observation suggests that block C residues may play a role in forming the active site, and perhaps contribute to the reaction mechanism.

11.5 AtEN101 as a functional marker

11.5.1 GUS activity and *PKT1*/*PKT2* expression

The isolation of several fusion transcripts overlapping the left junction of the T-DNA and the co-localization of their 5' ends with those of *PKT1* transcripts (Section 6.2) indicates that the promoter of *PKT1* is responsible for expression of the associated *gus*. The expression of *PKT1* in all organs tested (Chapter 8) further indicates that its promoter drives most (if not all) *gus* expression in the same organs of AtEN101. Nevertheless, the peculiar organization of the T-DNA locus in AtEN101 places some difficulties in establishing a relation between GUS activity and the expression of the *PKT1* and *PKT2* genes. One main difficulty derives from the presence of an apparently intact *gus* gene closely upstream from *PKT2* (Chapter 6). Although no evidence for expression of this gene was obtained, it could not be ruled out that it may be active. Particularly relevant in this respect was the observed functioning of the promoter region of *PKT1* in reverse orientation in transient assays, although the construct tested included an additional 0.2 kb not present downstream from the T-DNA in the transgenic line (Section 8.5.3.2).

Taking in consideration the above caveat and the problems derived from the stability of GUS (Section 11.2.4), it is not possible to infer the possible specific roles of *PKT1* and *PKT2* in AtEN101, or even the significance of the particular pattern of *gus* expression in the mutant. For example, the possible relation between the switching on and off of GUS activity in mutant embryos and the expression of *PKT1*/*PKT2* will have to be investigated by the specific study of the latter.

AtEN101 can nevertheless be a useful functional marker for the study of the regulation of *PKT1*/*PKT2*. As the structure of the tagged region indicates, GUS expression is dependent on the promoter of *PKT2* and, perhaps, on elements that may participate of the promoter of *PKT1*. Hence, alterations in the normal pattern of expression of GUS are likely to reflect changes in the regulation of either (or both) sets of elements. Once conditions are identified that alter the expression of GUS in AtEN101, their effects on the expression of the *PKT1*/*PKT2* genes can be tested. The advantage offered by this strategy resides in the easier screening for those

conditions. The observed effect of ABA on *gus* expression in seedling roots (Section 9.3.2) constitutes one example to be explored. Coupling the ABA-treatment with transcription analysis (e.g. Northern blot hybridization, quantitative RT-PCR and/or hybridization to cDNA arrays) should establish its possible effects on the expression of the *PKT1/PKT2* genes.

The repressing effect of ABA on GUS activity observed in seedling roots (Section 9.3.2) could have resulted from a general effect on the physiological condition of the plants, rather than being specific for the molecular marker. A possible test to distinguish the two possibilities could be performed by submitting seedlings of another transgenic line constitutively expressing *gus* to a similar parallel treatment. However, this was thought to be unnecessary as empirical observation failed to suggest that the ABA treatment slowed the development of the seedlings relative to untreated controls. Also noteworthy was the observation that GUS activity was reduced after a relatively short period of 24 h of exposure to 10 μ M ABA. This latter observation is also important in the context of the long half-life of the enzyme (see Section 11.2.4) suggesting that the expression of the enzyme was affected soon after the addition of the growth substance. Interestingly, while the repressing effect of ABA mimicked its expected role during the seed maturation period when the metabolic emphasis is on fatty acid synthesis rather than degradation, the apparently alleviating effect of sucrose was unexpected. This observation highlights one of the limitations of the use of promoter-traps as functional markers and the need to verify possible regulatory effects of test conditions on the native genes.

11.5.2 GUS activity in embryo-defective mutants

The utilization of AtEN101 as a functional marker in *hydra* and *gnom* mutant backgrounds (Section 9.4) showed that its pattern of expression is not altered in cotyledonary stage embryos. Both polar distribution and temporal control of activity were observed. This indicates that *HYDRA* and *GNOM* do not affect the genetic program controlling *gus* in AtEN101. The polar expression of GUS observed even in ball-shaped mutant embryos demonstrates the presence of biochemically distinct groups of cells in an otherwise apparently homogeneous population. These results were in good agreement with similar findings by Topping & Lindsey (1997) relative to the expression of other functional markers in *hydra* and *gnom* backgrounds, and their assertion that biochemical differentiation can take place independently of patterning and morphogenesis. The study illustrates the applicability of functional markers obtained by promoter-trapping in the study of developmental phenomena such as embryogenesis.

11.6 *bashful*: a new shape mutant

11.6.1 Non-genetic nature of morphological aberrations in *bashful* embryos

The screening of the library of *A. thaliana* lines transformed with p Δ gusBin19 (Clarke *et al.*, 1992) for mutants affected in seedling pattern or shape resulted in the identification of the new dwarf mutant *bashful*. As indicated earlier (Section 1.4.2 and 3.3.1), this screening strategy was aimed at the isolation of mutations affecting embryo development. Examination of embryos in homozygous plants showed a great diversity of shape and sizes (*e.g.* Fig. 3.6B). However, unlike in other mutants such as *gnom* and *hydra* where defective embryos also display a range of defects (Lindsey *et al.*, 1996), the morphological aberrations in *bashful* embryos do not seem to be caused by expression of their genome. There are two lines of evidence supporting this assertion. First, all seeds from heterozygous *bashful* plants have a normal appearance. Second, in spite of the great diversity of size and shape of the homozygous embryos observed, morphologically, *bashful* seedlings were rather uniform in shape. As indicated in Section 3.3.2, these observations together with the crumpled conditions found in the very reduced siliques of the mutant indicate that the aberrations are caused by physical constraints imposed on the developing embryo. It is however possible that the mutant embryos have other less obvious defects, in particular alterations of cell shape/number (see below).

11.6.2 The nature of the affected function in *bashful*

The morphology of the homozygous *bashful* plants throughout their life cycle was found to be very similar to those of other dwarf mutants including *dwf1/dim* (Kauschmann *et al.*, 1996; Takahashi *et al.*, 1995), *cpd* (Szekeres *et al.*, 1996), *det2* (Chory *et al.*, 1991) and *dwf4* (Azpiroz *et al.*, 1998). They are also similar in their developmental response to germination in the dark and different light conditions. In common with the other dwarf mutants, the cellular dimensions were affected in *bashful*, at least for those tissues that were examined. In the hypocotyl of dark grown seedlings and in roots, epidermal and cortical cells were smaller along the apical-basal axis than in the wildtype. Also in the hypocotyl of etiolated seedlings, the same cells were found to be expanded in the other directions. Whether other cells in other tissues, etc. (*e.g.* in embryos) are also affected is not known at the present time.

The similarity of the phenotypic traits to those of other dwarf mutants suggests that the mutation in *bashful* may affect a similar developmental or regulatory pathway, or a common shared step between pathways. Dwarfism can be caused by defects in gibberellin and brassinosteroid synthesis and perception (*e.g.* Azpiroz *et al.*, 1998). *dim*, *cpd*, *det2* and *dwf4* have been proposed to participate in a signalling pathway through their direct involvement in the biosynthesis of brassinosteroids with *CPD* and *DWF4* encoding cytochrome P450s that appear to be steroid hydroxylases (Szekeres *et al.*, 1996; Choe *et al.*, 1998). *DET2* specifies a steroid hormone 5 α -reductase (Li *et al.*, 1997), and, recently, *DIM* was found to encode a

cytochrome P450 monooxygenase involved in steroid synthesis (Klahre *et al.*, 1998). In this respect the observation that like *dim*, *cpe*, *det2*, and *dwf4* *bashful* seedlings respond to the presence of epibrassinolide in the growth medium suggests that it may also be affected in brassinosteroid synthesis.

Other interpretations of the preliminary results obtained with exposure to growth control substances are possible. Wildtype plants also respond with increased cell elongation to the presence of epibrassinolides in the media (*e.g.* Kauschmann *et al.*, 1996). It is then possible that the observed elongation of petioles and seedling growth of *bashful* plants exposed to epibrassinolide does not denote a lack of brassinosteroids in the plants but a normal response to increase levels of these substances. On the other hand, the apparent lack of response to the addition of gibberellic acid could be due to a defect in the perception of gibberellins. However, *dwf4* also fails to respond to concentrations of gibberellic acid up to 10^{-5} M (Azpiroz *et al.*, 1998), and *bashful* homozygous seeds do not require added gibberellins for germination. Furthermore gibberellins failed to recover the phenotype, as they did also with the other mutants.

11.6.3 A strategy to isolate *BASHFUL*

It follows from the above that the limited data available does not yet permit to determine the nature of the function affected in *bashful*. The isolation and characterization of the corresponding gene could provide important clues as to its role and helping interpret the phenotype exhibited by the mutant. Segregation analysis allowed the isolation of three families of the mutant carrying a single kanamycin resistance gene to which the phenotype was tightly linked (Section 3.3.6). Thus the affected function maps to or nearby the kanamycin resistance gene. The presence of a single functional kanamycin gene and of two apparently defective *gus* genes detected by Southern blot hybridization (Section 3.3) suggests that they are all part of a single if complex T-DNA present in the genome. It follows then that isolating and characterizing the tagged region of the genome is a straightforward proposition. Candidate genes for the defective function may be revealed by examination of their structures and their roles tested by transformation and complementation.

11.7 Future work

The characterization of the two lines, AtEN101 and *bashful*, that constituted the present work is a solid foundation for further studies concerning the nature and functional relevance of the tagged genes and their products.

11.7.1 *bashful*

It is of great interest to identify the mutated function in this line as it is clearly hormone-related and has a dramatic effect on plant development. The linkage to the T-DNA will be exploited to isolate the tagged region of the genome. Its molecular characterization may allow for identification of the mutated gene and its function. Other possible directions of study include, for example, testing the effects of various metabolites of the brassinolide biosynthesis pathway on development of homozygous *bashful* plants, to determine if the mutation affects one of its steps. The genetic characterization of the mutant by complementation studies with other dwarf mutants, although slower may prove fruitful for the identification of the affected function and the recognition of possible hierarchical relations between the various loci. Subsequent work on this line will depend on the results obtained with the initial studies.

11.7.2 *pAgusBin19*

It is of interest to determine if the potential 3' splice sites at -276 and -700 in the left border region are functional. This can be tested by RT-PCR in AtEN101 as well as other GUS-expressing lines which are currently being characterized. It would also be of great interest to determine if the region containing the first methionine codon of *gus* contains an internal ribosomal entry site.

11.7.3 *PKT* genes

The unexpected existence of a family of peroxisomal thiolases raises a number of questions relating to their specific metabolic significance and regulation. It is now possible to explore these in a variety of ways:

- First, the patterns of expression can be analyzed at the organ and stage levels by Northern blot hybridization and (quantitative) RT-PCR. However, the new technique of hybridization to cDNA arrays should be particularly powerful in the simultaneous analysis of the expression of all *PKT* genes as well as other genes involved in the same and related metabolic pathways. The results obtained may then be further refined, if necessary, by *in situ* hybridization.
- Purified PKT proteins can be obtained by expression of their cDNAs in *E. coli*. These then can be used to test their activities with potential substrates. Observed preferences may prove to be useful in establishing the functional relevance of the various isozymes and their genes.
- The presence and expression of *PKT1/PKT2* in the Landsberg cultivar should be examined to evaluate the significance of the apparent lack of complementation of *pkt3* (*ped1*) in this genomic background. Depending on the results, overexpression of *PKT2* in *ped1* may be

attempted to establish if the lack of complementation is due to a low level of expression of the gene or differences in substrate specificity (although this may also be tested biochemically, see above).

- *pkt4* mutants can be sought in interposon libraries by pool-PCR techniques. Analysis of double and triple mutants between AtEN101, *pkt4* and *ped1* (*pkt3*), should shed light on the possible overlapping and specificity of their functions.
- Re-transformation of *Arabidopsis* with the *PKT1* (and possibly the *PKT2*) promoter(s) driving a reporter gene should enable a less-hindered observation of its regulation.
- Furthermore, provided that the wave-pattern of expression in developing embryos is confirmed, the relevant regulatory *cis*-acting elements, and interacting *trans*-acting factors can be identified and isolated. Both would be useful in the study of regulation of biochemical differentiation during embryogenesis.
- The existence of a family of thiolase genes poses interesting questions from the point of view of their evolution. These can be addressed in several ways. As a first step, the possible existence of *PKT1/PKT2* equivalents in other species with varying degrees of relatedness to *Arabidopsis* will be explored. The molecular characterization of the possible fifth thiolase gene (Chapter 10) will also be pursued to determine its identity.

Appendix 1

Alignment of the insert sequence in pIPCR2 with the complementary sequence of the genomic fragment B1.4-E1.5. The alignment was produced with SeqAid™.

```

Comparison of pIPCR2          1 to   1095
with B1.4-E1.5              1 to   1092

      v      v      v      v      v      v      v
                                PIPCR2    70
GGATCCAAAGATGTTCTTTCCACTACAGCCTAATCAGAACAAACAATGAAACATAAGAAATCATTGATAC
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
GGATCCAAAGATGTTCTTTCCACTACAGCCTAATCAGAACAAACAATGAAACATAAGAAATCATTGATAC
                                B1.4-E1.5 70

      v      v      v      v      v      v      v
                                PIPCR2    140
TACTACTCCATGCATGCACAAGGTTTAATTGATGACTTACCTTAAGAACAGAAGCAAGAAGATCATCAGG
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
TACTACTCCATGCATGCACAAGGTTTAATTGATGACTTACCTTAAGAACAGAAGCAAGAAGATCATCAGG
                                B1.4-E1.5 140

      v      v      v      v      v      v      v
                                PIPCR2    210
AAGAGTGTCTTTGAAACCTCCACGTCTCGCTTTGCAAATGGCGGTACGATATGCCCTGAAGAAGAAGAAG
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
AAGAGTGTCTTTGAAACCTCCACGTCTCGCTTTGCAAATGGCGGTACGATATGCCCTGAAGAAGAAGAAG
                                B1.4-E1.5 210

      v      v      v      v      v      v      v
                                PIPCR2    280
AATCAAGAAACAAAGAAAACCTAAAGAAAGTGAAAGTGAAGCAAACCTCAATGAGTATACTTACGCTACAAT
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
AATCAAGAAACAAAGAAAACCTAAAGAAAGTGAAAGTGAAGCAAACCTCAATGAGTATACTTACGCTACAAT
                                B1.4-E1.5 280

      v      v      v      v      v      v      v
                                PIPCR2    350
CACAAATGTCATCTCCAAAAGCAGCCATTGGGGAAACTTCAGAAACACAATTCACAGGCTAAAATTAACAA
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
CACAAATGTCATCTCCAAAAGCAGCCATTGGGGAAACTTCAGAAACACAATTCACAGGCTAAAATTAACAA
                                B1.4-E1.5 350

      v      v      v      v      v      v      v
                                PIPCR2    420
AAGCACAAGTGAGTCACATAATAAAACAGAGTAAAAGATCATCAGCTTCCACAAACAATTTCAATGTCTA
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
AAGCACAAGTGAGTCACATAATAAAACAGAGTTAAAGATCATCAGCTTCCACAAACAATTTCAATGTCTA
                                B1.4-E1.5 420

      v      v      v      v      v      v      v
                                PIPCR2    490
ATTAAGACTCACGGATTAAGACAACCTAACTTACAGACAGAAGAGAAGGTTTCATGTTTAAGAGAAGAATT
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
ATTAAGACTCACGGATTAAGACAACCTAACTTACAGACAGAAGAGAAGGTTTCATGTTTAAGAGAAGAATT
                                B1.4-E1.5 490

      v      v      v      v      v      v      v
                                PIPCR2    560
AGAAGAAGAAACTGGATTGAGATGACGAAGCAATATCTTTTGTCTTTCCATAGCTCTCTCCATGATTCTC
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
AGAAGAAGAAACTGGATTGAGATGACGAAGCAATATCTTTTGTCTTTCCATAGCTCTCTCCATGATTCTC
                                B1.4-E1.5 560

```

```

      v      v      v      v      v      v      v
      PIPCR2      630
TCTGTTGCAAACACTTGTGTTGCAGCTGAATTCTAATGGTGATTGAAAGACAAAACTTTCCGGGGAAAC
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
TCTGTTGCAAACACTTGTGTTGCAGCTGAATTCTAATGGTGATTGAAAGACAAAACTTTCCGGGGAAAC
      B1.4-E1.5 630

      v      v      v      v      v      v      v
      PIPCR2      700
TGAAATGGGTTTGGTTGATTTTTCTGGAAAGGGGAAGACTTTAAGTACAATGAGCTTCAGGGTCCTCATT
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
TGAAATGGGTTTGGTTGATTTTTCTGGAAAGGGGAAGACGTTAAGTACAATGAGCTTCAGGGTCCTCATT
      B1.4-E1.5 700

      v      v      v      v      v      v      v
      PIPCR2      770
TTTATCGTGGGAGAGATTTTTGATCGATCGTTCTCAGTTCTCACAAGTCTGGTCTGATTGGCCTCGTTGC
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
TTTATCGTGGGAGAGATTTTTGATCGATCGTTCTCAGTTCTCACAAGTCTGGTCTGATTGGCCTCGTTGC
      B1.4-E1.5 770

      v      v      v      v      v      v      v
      PIPCR2      840
CAAAACCTTGTCTTGTCTGATTTTGTATAAGTTAACGTCGTTACGAAAATATCGCATTTTTGACATGTTTT
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
CAAAACCTTGTCTTGTCTGATTTTGTATAAGTTAACGTCGTTACGAAAATATCGCAT TTTGACATGTTTT
      B1.4-E1.5 839

      v      v      v      v      v      v      v
      PIPCR2      910
ACGCCACATACAAATCACGCCGTTAAAAATGTGACCGGGAATTAATCGTCGGCAACTATAACTAAACCGG
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
ACGCCACATACAAATCACGCCGTTAAAAATGTGACCGGGAATTAATCGTCGGCAACTATAACTAAACCGG
      B1.4-E1.5 909

      v      v      v      v      v      v      v
      PIPCR2      980
GCCACGTGGGAAGAATCTAACTTTTAAGCAATAAATGTTTCTTCTCATTTTGGGTATTAGTGATTTTGAT
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
GCCACGTGGGAAGAATCTAACTTTTAAGCAATAAATGTTTCTTCTCATTTTGGGTATTAGTGATTTTGAT
      B1.4-E1.5 979

      v      v      v      v      v      v      v
      PIPCR2      1042
AATTTGAGTAATTTA      GAAGAATTTGAGATAATTTATAAAATTTTAGTGAATAATTTGGATAT
::::::::::::::::::::: :::::::::::::::::::::::::::::::::::::::::::::::::::::
AATTTGAGTAATTTACAAATTTAGAAGAATTTGAGATAATTTATAAAATTTTAGTGAATAATTTGGATAT
      B1.4-E1.5 1049

      v      v      v      v      v      v      v
      PIPCR2      1095
CCAGATAATATACATGATTTGTATTCTTTACAGATAATTACAGATAATTTCAA
:::::::::::::::::::::::::::::::::::: : :: :
CCAGATAATATACATGATTTGTATTCTTTACCCACAAGAATTC
      B1.4-E1.5 1092

```

Total Number of Matches: 1073

Sequencing chromatogram of 5' RACE plasmid pTM-7 (Section 5.4.3).



Appendix 3

Alignment of the PLFR-7 and pΔ*gus*Bin19 sequences.

The alignment was performed with SeqAid™.

Comparison of PLFR-7 with pΔ <i>gus</i> B19	1 13350	to	547 13911			
v	v	v	v	v	v	v
					PLFR-7	63
GGGGGGGGGGGGGGTNCAGTTGTTTCCTTACTGGGCTTCCTCAGCCCCAGATCTGGGGTCGAT						
:	:	:	:	:	:	:
.....TACAGTTGTTTCCTTACTGGGCTTCTCAGCCCCAGATCTGGGGTCGAT						
					PDGUSB19.SEQ	13419
v	v	v	v	v	v	v
					PLFR-7	133
CAGCCGGGGATGCATCAGGCCGACAGTCGGAACTTCGAGTCCCCGACCTGTACCATTCCGTGAGCAATGG						
:	:	:	:	:	:	:
CAGCCGGGGATGCATCAGGCCGACAGTCGGAACTTCGGGTCCCCGACCTGTACCATTCCGTGAGCAATGG						
					PDGUSB19.SEQ	13489
v	v	v	v	v	v	v
					PLFR-7	202
ATAGGGGAGTTGATATCGTCAACGTTCACTTCTAAAGAAATAGCGCCTCTCAGCTT					CTCAGCGGCTTTTA	
:	:	:	:	:	:	:
ATAGGGGAGTTGATATCGTCAACGTTCACTTCTAAAGAAATAGCGCCACTCAGCTTCCTCAGCGGCTTTA						
					PDGUSB19.SEQ	13559
v	v	v	v	v	v	v
					PLFR-7	272
TCCAGCGATTTCTATTATGTTTCGGCATAGTTCTCAAGATCGACAGCCAGTCACGGTTAAGCGAGAGATG						
:	:	:	:	:	:	:
TCCAGCGATTTCTATTATG	TCGGCATAGTTCTCAAGATCGACAGCCTGTACAGGTTAAGCGAGAAATG					
					PDGUSB19.SEQ	13628
v	v	v	v	v	v	v
					PLFR-7	317
AATTA	TGCGAG	GAGATGTATGT	ATCACAGGCAGCAACGCTCTGTCT			
:::	:	:	:	:	:	:
AATAAGAAGGCTGATAATTCGGATCTCTGCGAGGGAGATGATATTTGATCACAGGCAGCAACGCTCTGTCT						
					PDGUSB19.SEQ	13698
v	v	v	v	v	v	v
					PLFR-7	387
ATCGTTACAATCAACATGCTACCCTCCGCGAGATCATCCGTGTTTCAAACCCGGCAGCTTAGTTGCCGTT						
:	:	:	:	:	:	:
ATCGTTACAATCAACATGCTACCCTCCGCGAGATCATCCGTGTTTCAAACCCGGCAGCTTAGTTGCCGTT						
					PDGUSB19.SEQ	13768
v	v	v	v	v	v	v
CTTCCGAATAGCATCGGTAACATGAGCAAAGTCTGCCGCC						
:	:	:	:	:	:	:
CTTCCGAATAGCATCGGTAACATGAGCAAAGTCTGCCGCC						

Appendix 4

List of deoxyoligonucleotides used in various experiments described in the text.

LB-169	CCAGCGTGGACCGCTTGCTGCAAC
LB-267	CAAACTCAACCCTATCTCGGGC
LB-63	CGTCCGCAATGTGTTATTAAG
Nested2	CAGGACGTAACATAAGGG
prANCHOR	CTGGTTCGGCCACCTCTGAAGGTTCCAGAATCGATAG
prEN1-P1	AACCTCCACGTCTCGCTTTGC
prEN1-P2	CGCTACAATCACAATGTCACTCTCC
prEN1-P3	GCAAGAAGATCATCAGGAAGAGTGTC
prGUS-3'	GCCGCCGACTTCGGTTTGCGGTG
prGUS-4247	TGAACAACGAACTGAACTGG
prGUS-5'	CTAACGCGCTTTCCACC
prGUS-cDNA	TGACCGCATCGAAACGCAGCACG
prGUS-CONTROL	CCCTTATGTTACGTCCTGTAGAAACCCC
prGUS-PCR	CAGGACGTAACATAAGGGACTGACCAC
prMYC-F	GAACAAAAGCTTATTTCTGAAGAGGATCTTAAGT
prMYC-R	AATTCAGTTAAGATCCTCTTCAGAAATAAGCTTTTGTCTGCA
prTHIO-1R	ACACGAGCGTTGGACAAGTTATC
prTHIO-2F	AAGATATGGCTGCGGTGGAGTCTC
prTHIO-3F	CTCTTACTAGCTCATTCATCCG
prTHIO-3R	CACTGATCTGACTAGCATTGCCAGC
prTHIO-4F	CAGCTGCAACACAAGTGTTC
prTHIO-4R	CAGTTGACATTGATTCCACTCC
prTHIO-5R	GCACCAATACCAATGTCGTAA
prTHIOP-F	GTTTGGCTATGAAGAAGGG
prTHIOP-R	AGTACACATACTGAGATGC
prTN δ	AGGGGAAGTGAAGAGCTCTA
prTN γ	TCAATAAGTTATACCAT
prTN1000-RPT	CGGGATCCTTCGTTCCATTGGCCCTCAAACCCC
prUAP	CTACTACTACTAGGCCACGCGTCGACTAGTAC

Appendix 5

Alignment of the EST T75842 with the mP18 sequence.

The alignment was performed with SeqAid™.

Comparison of T75842.EST: 1 to 367

with mP18: 1 to 400

```

      v      v      v      v      v      v      v
                                T75842.EST  44
                                ATCTATTGGAACCTCTTACTAGCTCATTTCATCCGTCGGCCTGTGA
                                ::::::::::::::::::::::::::::::::::::::::::::::::::::
CATATTGTCTACTTTTAGATTTTGATATCTATTGGAACCTCTTACTAGCTCATTTCATCCGTCGGCCTGTGA
                                mP18      70

      v      v      v      v      v      v      v
                                T75842.EST  114
ATTGTGTTTCTNAAGTTTCCCAATGGCTGCTTTTGGAGATGACATTGTGATTGTAGCGGCATATCGTAC
:::::::::::: ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
ATTGTGTTTCTGAAGTTTCCCAATGGCTGCTTTTGGAGATGACATTGTGATTGTAGCGGCATATCGTAC
                                mP18      140

      v      v      v      v      v      v      v
                                T75842.EST  184
CGCCATTTGCAAAGCGAGACGTGGAGGTTTCAAAGACACTCTTCCTGATGATCTTCTTGCTTCTGGTNCT
:::::::::::: :::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::: : ::
CGCCATTTGCAAAGCGAGACGTGGAGGTTTCAAAGACACTCTTCCTGATGATCTTCTTGCTTCTGTT CT
                                mP18      209

      v      v      v      v      v      v      v
                                T75842.EST  254
TAAGGCTGTAGTGGAAGAACATCTTTGGATCCAAGTNAAGTTGGTGATATCGGTTGTTGGGACCGTGNA
:::::::::::: :::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::: : :
TAAGGCTGTAGTGGAAGAACATCTTTGGATCCAAGTGAAGTTGGTGATATC GTTGTGTTGTTACCGTG A
                                mP18      277

      v      v      v      v      v      v      v
                                T75842.EST  324
TAGCGCCTNGGTTCTCAGAGAGCAATGGGAGTTTAGAGTTNCAGCTTATTTTGGCCTGGGTTTNCCTNA
:::::::::::: :::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::: : :
TAGCGCCT GGTTCAGAGAGCAATGGAGTG TAGAGTTGCAGCTTATTTTGCTGGTTCCTGACTCC
                                mP18      345

      v      v      v      v      v      v      v
                                T75842.EST  367
CTTCCGGGNCCANTTAGGAACCTTNTCAATNNGGGCAANTGCC
: : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
GTGCCAGTTAGAACTGTCAATAGACAATGCTCATCAGGACTACAAGCAGTTGCTG
                                mP18      400

```

Total Number of Matches: 305

Appendix 6

Alignment of the EST AI100014 with the mP19 sequence.
The alignment was performed with SeqAid™.

Comparison of AI100014: 1 to 393

with mP19: 1298 to 1707

```

      v      v      v      v      v      v      v
                                AI100014      68
CCTGCAAGAAGTTAGACTN GATATGGAAAAGN CAATGTTAATGGAGGANCNATNGCTATTGCCNACCC
: ::::::::::::::: ::::::::::::::: ::::::::::::::: : : ::::::::::: : : ::
CTTGCAAGAAGTTAGAGCTGGATATGGAAAAGGTCAATGTTAATGGAGGAGCCATTGCTATTGGCCATCC
                                mP19      1367

      v      v      v      v      v      v      v
                                AI100014      137
CCTGNGTCTTANAGNAGCTCGATGTGTTGCGACATT TTGCACGAGATGAAGCGGAGAGGAAAAAGATTG
:::: : : : : ::::::::::::::: ::::::::::::::: :::::::::::
CCTGGGTGCTACAGGAGCTCGATGTGTTGCGACATTGTTGCACGAGATGAAGCGGAGAGGA AAAGATTG
                                mP19      1436

      v      v      v      v      v      v      v
                                AI100014      207
CCCGCTTTGGAGTAATCTCAATGTGCATAGGCACTGGAATGGGAGCTGCAGCTGTTTTGAGAGGGGAGA
::::::::::::::::::::: ::::::::::::::: :::::::::::::::
CCGCTTTGGAGTAATCTCAATGTGCATAGGCACTGGAATGGGAGCTGCAGCTGTTTTGAGAGGGGAGA
                                mP19      1505

      v      v      v      v      v      v      v
                                AI100014      277
CTCTGTTGATAACTTGTCCAACGCTCGTGTGGCTAACGGGGATAGTCATTAGAACATCGAAGAGAGCTTG
::::::::::::::::::::: ::::::::::::::: :::::::::::::::
CTCTGTTGATAACTTGTCCAACGCTCGTGTGGCTAACGGGGATAGTCATTAGAACATCGAAGAGAGCTTG
                                mP19      1575

      v      v      v      v      v      v      v
                                AI100014      347
AATAAGTAGAAGTAATGATGCATTGAGTCTNATAAATATGATGCTTTAGCTCTTTCACATTGCTGAACAA
::::::::::::::::::::: ::::::::::::::: :::::::::::::::
AATAAGTAGAAGTAATGATGCATTGAGTCTAATAAATATGATGCTTTAGCTCTTTCACATTGCTGAACAA
                                mP19      1645

      v      v      v      v      v      v      v
                                AI100014      393
TGAAAACTTTTGTCATTCTGAGTTTAAATCAACTACTTTTCTCTG
::::::::::::::::::::: :::::::::::::::
TGAAAACTTTTGTCATTCTGAGTTTAAATCAACTACTTTTCTCTGAAAAAAAAAAAAAAAAAAAA
                                mP19      1707

```

Total Number of Matches: 374

Appendix 7

Multiple alignment of the EST sequences encoding ORFs with strong similarity those encoded by the (putative) *PKT* genes.

The four EST sequences, accession numbers AA404882, H36507, T45052, T76117 were aligned with PILEUP using a gap weight and gap extension weight penalties of 5.0 and 0.3, respectively. A consensus sequence (**Consens.**) was derived from the segment where it was reliable to do so. A translation of the consensus in its frame 2 shows strong similarity to the (putative) PKT proteins (see Section 10.3.2; Fig. 10.6).

AA404882	1	GGCTCTTCAG	CTAGGACTTC	TAGTATTAGC	AAAAATTAAA	GGGTATGGTG	ACGCAGCTCA	GGAACCAGAG	70
H36507		GGCTCTTCAG	CTAGGACTTC	TAGTATTAGC	AAAAATTAAA	GGGTATGGTG	ACGCAGCTCA	GGAACCAGAG	
T45052		
T76117		
Consens.		GGCTCTTCAG	CTAGGACTTC	TAGTATTAGC	AAAAATTAAA	GGGTATGGTG	ACGCAGCTCA	GGAACCAGAG	
AA404882	71	TTTTTCACTA	CTGCTCCTGC	TCTTGCTATA	CCAAAAGCCA	TTGCACATGC	TGGTTTGGAA	TCTNCTCAAG	140
H36507		TTTTTNACTA	CTGCTCCTGC	TCTTGCTATA	CCAAAAGCCA	TTGCACATGC	TGGTTTGGAA	TCTTCTCAAG	
T45052	CCNGA	CCGNTTTCNN	CTGCCNCTN	TTNCCAAANC	NNTCNCNTCT	NGTTGGANCT	
T76117		CCTAAAGCTA	TAAAGCGGCG	TGGTTTGGAT	GCATCTCAAG	
Consens.		TTTTTCACTA	CTGCTCCTGC	TCTTGCTATA	CCAAAAGCCA	TTGCACATGC	TGGTTTGGAA	TCTTCTCAAG	
AA404882	141	TTGATTACTA	TGAGATCAAT	GAAGCATTTC	CAGTTGTAGC	ACTTGCAAAT	CAAAAGCTAC	TCGGGATTGC	210
H36507		TTGATTACTA	TGAGATCAAT	GAAGCATTTC	CAGTTGTAGC	ACTTGCAAAT	CAAAAGCTAC	TCGGGATTGC	
T45052		CCAATTGATT	CTATGNNNCA	ATGAANCATT	GCATTGTANC	ACTTGCAA..	TCAAAANTAC	TCGGGA.TGC	
T76117		TGGATTATTA	TGAAATAAAC	GAAGCATTTC	NTGTGGTAGC	TCTAGCCAAT	CAGAACTAC	TGGGATTAGA	
Consens.		TTGATTACTA	TGAGATCAAT	GAAGCATTTC	CAGTTGTAGC	ACTTGCAAAT	CAAAAGCTAC	TCGGGATTGC	
AA404882	211	TCCAGAGAAA	GTGAACGTAA	ATGGAGGAGC	TGTCTCCTTA	GGACACCCCTC	TAGGCTGCAG	TGGCGCCCGT	280
H36507		TCCAGAGAAA	GTGAACGTAA	ATGGAGGAGC	TGTCTCCTTA	GGACACCCCTC	TAGGCTGCAG	TGGCGNCCGT	
T45052		CCCAGAGAAA	GTGAACGTAA	ATGGAGGAGC	TGTCTCCTTA	GGACACCCCTC	TAGGCTGCAG	TGGCGCCCGT	
T76117		TCCTGANCGG	CTCAATGCGC	ATGGAGGGGC	TGTTTCACTG	GGACATCCAT	TGGGCTGTAG	CGGTGCTCGT	
Consens.		TCCAGAGAAA	GTGAACGTAA	ATGGAGGAGC	TGTCTCCTTA	GGACACCCCTC	TAGGCTGCAG	TGGCGCCCGT	
AA404882	281	ATTCTAATCA	CGTTGCTTGG	GATACTAAAG	AAGAGAAACG	GAAAGTACGG	..TGTNGGAG	GAGTGTNCAA	350
H36507		ATTCTAATCA	CGNTGCTTGG	GNTACTAAAG	AAGNGGAACG	GGAAGTNCGG	GTNTTGGGAG	GAGT'TTTCAA	
T45052		ATTCTAATCA	CGTTGCTTGG	GATACTAAAG	AAGAGAAACG	GAAAGTACGG	..TGTGGGAG	GAGTGTGCAA	
T76117		ATCTTGCTCA	CATTATTTGG	GGTGTGAGA	GCAAAGAAGG	GAAAGTATGG	..AGTGGCAT	CAATATGCAA	
Consens.		ATTCTAATCA	CGTTGCTTGG	GATACTAAAG	AAGAGAAACG	GAAAGTACGG	..TGTGGGAG	GAGTGTGCAA	
AA404882	351	CGNA.GGAGG	AGGTGCTTCT	GCTCTAAGT.	TCTTAGCNCC	NTTGATGCAT	TATGAATCCA	NGGTTGGTAA	420
H36507		GGGAGGNAGG	AGGTGCCTTT	NCTCTAGTTC	NTTTNGCNCC	CTTNATNCAT	TTTTNATCCN	GGTTGGAAAC	
T45052		CGGA.GGAGG	AGGTGCTTCT	GCTCTAGTTC	TTGAG.CTCC	TTTGATGCAT	TATGAATCCA	GGTTGGTAAA	
T76117		CGGA.GGAGG	AGGAGCATCA	GCACTTGTC	TTGAGTTCAT	GTCGGAGAAG	ACAATCGGNT	ATTCGGCACT	
Consens.		CGGA.GGAGG	AGGTGCTTCT	GCTCTAGTTC	TTGAG				
AA404882	421	CNNAAATAGN	ATGNTCCATC	TNTTCATNCA	ACTTCCATNG	NNCAGTCATT	TCACTTNNGC	TTTATTTNAG	490
H36507		
T45052		CTAAA..TAG	ATGGATCTAT	CTCTCATCCA	ACTTCCATGA	GTCAGTCATT	GCACTTTACG	TTTTATGTTA	
T76117		CTGAAGCTAA	GAGCTGTTGT	TTTATAACGC	AAAGCTTNTG	TATATGGTTC	ANGTGT'TNT	CACAAGNCTG	
AA404882	491	GTTGGCCCCC	AAAAAGGGGN	ACCCTTGAA.	560
H36507		
T45052		GTTGGTCTCT	CAATAAAGGA	GTGATCGTTG	ATGTTGTAA	GTCTATTTAT	TTATATGAAC	TCGGCTTCTT	
T76117		TAAACTTTGT	NCC'TT'NNGG	GGGTTNCAAT	AAAAGTTT'N	TTNG.....	
AA404882	561	612
H36507		
T45052		TTTTTCTCTT	ACCGGTGACA	GTCAGTTATT	GAATCATGGA	CCGATTTACC	GT		
T76117			

Appendix 8

Multiple alignment of *PKT2*, *PKT3* and *PKT4* sequences.

Alignment of the *PKT2*, *PKT2*, and *PKT3* sequences from their first methionine codons.

Exons are contained within the arrowheads below the alignment (< < > >). In the two instances where the exon limits differ, more the different limits are indicated above and below the alignment. The alignment was obtained with Pileup (gap penalty=3.0 and gap extension penalty=0.1). Conserved residues are indicated by the darker backgrounds generated with GeneDoc.

[illegible]

```

      700      *      720      *      740      *      760      *      780      *      800
PKT3 : ACATAGGACTCCACTATGCAAGTCCAAACGTGGCAATTTCAAGGATACATATCCCGATGATTTGCTCGCACCTGTTTGGAGGGTCTGTGCTATCTTTCTATTCTGTTGTTGGT : 805
PKT4 : ACAAGGACTGCACCTTGCAAGGCCAAACGTGGCAGCTTCAAGGATACATTTCCAGACGAGTTGCTTGCCTCTGTATTGAGAGTATGCTTGTTCAGTTCTGTTTCCTGTTT : 583
PKT2 : ATATCGTACCGCCATTTGCAAAGCGAGACGTGGAGGTTTCAAAGACACTCTTCCTGATGATCTTCTTGCTTCTGTTCTTTAAGTAAGTCATCAATTAACCTTGTGCATGCAATGG : 476
                                     3>>

      *      820      *      840      *      860      *      880      *      900      *      920
PKT3 : TTTTATTTCTTTTCTAGGTTCTATTATAGATAGTTCTTCTTGACATAGGCATTGATAGAGAAAGACGAATCTAAACCCAAGTGAAGTAGGTGACATTGTTGTGGGTACTGTTTGG : 920
PKT4 : TGGCTTCCCTTTTTCCTAATTTGCACTTATGTGTGCTTCTTCTTGCTAGGCATTGATAGAGAAAACATAATGTAACCCAAGTGAAGTTGGTGACATTGTAGTGGGTACTGTTTGG : 698
PKT2 : AGTAGTAGTATCAAT-GATTTCTTATGTTTCTTTGTTTCTGATTAGGCTGTAGTTGAAAGAACATCTTTGGATCCAAGTGAAGTTGGTGATATCGTTGTTGGTACCGTGATA : 590
                                     <<4

      *      940      *      960      *      980      *      1000      *      1020      *
PKT3 : GCACCTGGATCTCAGAGAGCCAGTGAATGCAGGATGGCTGCCTTCTATGCTGGTTTCCTGGTATGTTTCTTCTATGTACAGTACAATACAATGTGAAAAGTCTTTGTTTTTAT : 1035
PKT4 : GGACCAGGATCTCAGAGAGCCAGTGAATGCAGGATGGCTGCCTTCTATGCTGGTTTCCCGGTAATATTTTTCTTCTGT--ATTATGGTAATCCTGAATCTTTTCTTCTTTTCTGT : 811
PKT2 : CGCCCTGGTCTCAGAGAGCAATGAGTGTAGAGTTGCAGCTTATTTTGCTGGTTTTCTTGGTACTAAATCTCATTCAGAATCAAGATAAACT---TAATATCTCACAAGTTGT : 702
                                     4>>

      1040      *      1060      *      1080      *      1100      *      1120      *      1140      *
PKT3 : TTCCTGTATGCTCATCCT-GTTATGTTTCAATTTACAGAAACCGTGGCTGTGAGAACGTGAATAGACAGTGGTCACTCTGGGCTTCAGGCTGTGCTGATGTAGCCGCTGCCATTAA : 1149
PKT4 : TTGTTGTATGCTTATGCTTCTTTTGTCTGTTGGAAGAAACTGTTCCCATCAGAACCGTGAACAGACAGTGTTCATCTGGGCTTCAGGCTGTGCTGATGTGCGCGCTGCCATAAA : 926
PKT2 : TGATATGTTTCTTCATCCTTCTCTTTTACGACTAGACTCGTGCCAGTTAGAACGTGCAATAGACAATGCTCATCAGGACTACAAGCAAGTTGCTGATGTGCTGCTTCCATTAG : 817
                                     <<5

      1160      *      1180      *      1200      *      1220      *      1240      *      1260
PKT3 : AGCGGGATTTTATGACATTGGTTAGTGAAT-----AACCCTGTTTAACTAG-----GCCACATTTTTCATAGAGAACCAGCTGACGATATGATGGAT---GCAGGTATCGG : 1249
PKT4 : AGCTGGTTTATTAGACATTGGTTAGATAGT-----AAGTATGTTTGGTTAGAAAAATTTCTTATCTTGGCATGAAGAATTGAGCTGAAATTAAGTTGAAT---ACAGGTATTGG : 1033
PKT2 : AGCTGGTTATTAGACATTGGTTAGATGCTGCTGCATCAAGTTTTTTTGTAGTCTCTAGCTGCTTGTATAGGCATGAGATTGATTCTCTCAACTTGTGAATGCATAGGTATTGG : 932
      5>>                                     <<6

      *      1280      *      1300      *      1320      6>>      *      1340      *      1360      *      1380
PKT3 : GGCTGCTTTGGAGTCCATGACTACCAATCCAATGGCATGGGAAGGGTCAGTCAACCCAGCGGTAC-----TGATTCTCTCTTATCTT-----GCACTCTTACTTGGCTC : 1347
PKT4 : AGCTGGGCTGGAGTCCATGACAACTAATCCAAGGGGATGGAAGGATCAGTCAACCCAAATGTACCTTGCTCTTATCTCTCTCTCTCTTCCCGAAGTGTATTGATTGGCTC : 1148
PKT2 : TGCTGGAGTGAATCAATGTCAACTGATCATAATCTTGAAGCGGCTTTCATGGCTCTAATCCAAGA-----GTAAGCTATCACAAGAGTGTAAATTTACTCAACA : 1036
                                     6>>

```



```

      *      1400      *      1420      *      1440      *      1460      *      1480      *
PKT3 : TGTATTATTTAAACGTTACTATATAACGGACTATGTTATTTTCAGGTGAAGAAGTTTGCACAAGCGCAGAAATTGTTCTTCTTCCTATGGGTGTTACGTCAGAAAATGTAGCACAAACGC : 1462
PKT4 : TGTGAATGATTAGCTCTCATAGAACGTTATTCGTTTGTGTTTAGGTGAAGAAGTTTGAACAAGCTCACAAATTGCCCTTCTTCCAATGGGTATTACTTCAGAAAATGTAGCACACCGG : 1263
PKT2 : AGAAGAGTTTGTGATGATCTTCTAAACTCTACATGGAATAGGCACAGGATTTCCAAAAGCCGTGATTGTTTGCTTCCAATGGGAATTACTTCTGAAAACGTTGCAGAAAGG : 1151
      <<7

      1500      *      1520      *      1540      *      1560      *      1580      *      1600      *
PKT3 : TTTGGTGTCTCAAGGCAGGAGCAAGATCAAGCTGCTGTAAGGGAACGAACTTTTTCAGTATGGTGTATTTTAAATAACCTATAATTTTGATCGCTATGCTAAAGGATGGAACTC : 1577
PKT4 : TTTAATGTTTCAAGGGAGGAGCAGGATCAAGCTGCTGTAAGTCAATAAGCAATCTTAGTTCTAATGCATGTT-----ATTTCGATATTGCCCTCCATCCCTTAGTTTATCGAATTC : 1372
PKT2 : TTCGGTGTCAACAAGAGAAGAGCAAGATATGGCTGCGGTTAGTTAGTTATTCTCTTCACTT-----TTTTCTCTCTAGTTCATATAAGTACTGATTCACTA : 1247
      7>>

      1620      *      1640      *      1660      *      1680      *      1700      *      1720
PKT3 : ATATTGCTAGGTTGACTCGCACAGAAAAGCAGCTGTGCTACTGCTGCTGGTAAATTCAAGGATGAGATCATTCTGTTAAGACCAAGGTA-TTTTCAGATCCTTCATACAATA : 1691
PKT4 : AAAATGTGTAGGTTGATTCTTCACAGAAAAGGCTGCTTCTGCTACTGCTTCCGGAAGTTTAAGGATGAGATAACCCCTGTAAAAACCAAGGTACATGCTGTTTCCACCTTCGAACA : 1487
PKT2 : ATAAAT--CAGGTGAGTCTCACAAACGCGCTGCAGCTGCAATCGGCTCTGGTAAACTCAAGGATGAATCATTCTGTGCTACTAAGGTA-----A : 1337
      <<8      8>>

      *      1740      *      1760      *      1780      *      1800      *      1820      *      1840
PKT3 : ACTGT-AACATTTAAGACTTCTCTTGGTGGGATTAGTTACTGAACTTCATTCATGTTCTAGTATTAGTCAATAACGTGTGTTTTTAAATTTTCAGCTGTGACCCGAAGACTGGTGA : 1805
PKT4 : ACTGTGAACATGTTAGATTCTTCAGGATGCATCAAGTGCT-----CTTACCCAAATTTAATGAATTTTGTGTTTACATTTTCAGATTGTTGACCCAAAGACAGGTGA : 1588
PKT2 : AACAAAGAACCTCCCAAACTCTTTTAAAGCCCTTTTGTAGT-----GCTTCTCAGTAAATGATATTGTTCTTGACAGATTGTGGACCCTGAGACTAAAGC : 1434
      <<9

      *      1860      *      1880      *      1900      *      1920      *      1940      *
PKT3 : TGAGAAACCCATTACAGTTTCTGTTGATGATGGTATCCGACCAACCCCAACTCTTGTCTTCTGTTGGGAAGCTGAAGCCAGTGTTTAAGAAGGATGGCACCCTACTGCTGGTACT : 1920
PKT4 : TGAGAAACCCATAACAGTTTCTGTGGATGATGGGATTCGACCTAACCAACCCCTTTCCGGAATTGCAAGCTGAAGCCAGTGTTTAAGGAAGAGCGAACCAACAACCTGCTGGTACT : 1703
PKT2 : AGAGAAGGCAATCGTCGTATCTGTTGATGACGGTGTACGTCCAACTCAAAATGGCAGATTGCGCAAAGCTGAAGACTGTCTTTAAACAGAACGGTTCACACAGCTGGTTAG : 1549
      9>>

      1960      *      1980      *      2000      *      2020      *      2040      *      2060      *
PKT3 : TTTCCCGAATCAATTGTTTCATTATGGAACACASTTCTTATATACAAATTTTAAATGCATTGCTTCATTTGTTCTGTGCGAAACAGGAATTCAGCCAAAGTAAGTGATGGT : 2035
PKT4 : TTTCCCGAGATTAAATTGGTCATTTCAGAGAAAATGTTTGTGTTGATC-CATCTTTTCCCAAAACCCCTGTTTCTTTGTTGTCAAACAGGGAATTCTAGCCAATTAAGTGACGGT : 1817
PKT2 : AAACAACTCTCTTAACTATCACTGGCATAAGTGAGATTATAGATTTGAAGCAATTAAGTATA-----TGTGTTTGTGTAGGCAATGCTAGTCAGATCAGTGATGGT : 1651
      <<10

```

```

      2080          *      2100          *      2120          *      2140          *      2160          *      2180
PKT3 : GCAGGAGCGTTCTCTCTAATGAAGAGAAGTGTTCGAATGCAGAAAGGACTTCCCGTTCTTGGTGTATTAGGTGGGTAATAATTACTTACAAAGCTACGACTGTGGAAGTCTGAA : 2150
PKT4 : GCTGGAGCTGTTCTCTCTATGAGGAGAAATGTCGCAATGCAGAAAGGCCCTTCCCATTCCTTGGTGTATTAGGTGGG----- : 1893
PKT2 : GCTGGAGCACTACTGCTAATGAAGAGAAGTTTGGCTATGAAGAAGGACTTCCCATTCCTTGGAGTATTAGG----- : 1723
                                     10>>

      *      2200          *      2220          *      2240          *      2260          *      2280          *      2300
PKT3 : AACCGAAAAAAGTGCAGAGGCTTGTGATTTCCATTACACAGATATGCTAGTATACGTGGAAAATACATGTTCTTCTGTATTTGGCTTTTCTTGTAGCACACTATTGGACA : 2265
PKT4 : -----CTTGATACCTTGATACTCTATATTGACCTTTCTTTGCTTCATTTAGTAACAGCACTCTTGGACA : 1956
PKT2 : -----TAAAATGAATATACACATTTTCCAGTTCAAT-----AAACCGCCCAATCTGATTCGGAA : 1781

      *      2320          *      2340          *      2360          *      2380          *      2400          *
PKT3 : TATTGCAGGACATTTGCTGCAGTTGGTGTGTGACCCGCAATCATGGGTATCGGTCCAGCACTTGCCATTCTTGCTGCAGTTAAGCGGGCTGGTTAGAACTTGATGACATGGACT : 2380
PKT4 : CATTACAGGACATTTTCTGCAGTTGGTGTGGACCCAGCCATCATGGGGGTTGGGCCAGCCGTTGCCATTCTTGCTGCAGTCAAGGCAGCTGGTTAGAACTCAATGACGTGGACT : 2071
PKT2 : CAATGCAGGAGTTTGTCTGTACTGTTGTGGAACCATCTGTAATGGGTATTTGGTCCAGCTGTTGCCATTCCCGCTGCAACTAAGCTCGCAGGGCTCAACGTCAGCGATATTGATC : 1896
      <<11

      2420          *      2440          *      2460          *      2480          *      2500          *      2520          *
PKT3 : TGTTTGAGATCAATGAGGTATATAGCAAAATTTTCATGATCTTCTTATCAATATCTC-----TTCCCGACATTCTGATGATGACCATCTGTTGAATCCT--TGTC : 2482
PKT4 : TGTTTGAGATTAAAGAGGTTAAGATAATAGAACAACTCTTCTTCTTTTAACTTTACTCAGCTCATAGTCTATTACTTATAGGCCTTATATTGTTTGTCTAAATCCTACTGTG : 2186
PKT2 : TATTTCGAGATCAATGAGGT-----AAGAAATATTTTCACATTGGTTTTCCTGT-----TTCGGAATAATACATACATATAAATCTATTGTTGCATTGT : 1989
      11>>

      2540          *      2560          *      2580          *      2600          *      2620          *      2640
PKT3 : TGTTTCAGGCATTTGCATCTCAGTTTGTTTATTGCGCTAACAAATTTGGGACTTGACCCAGAGAAAATCAATGTCAACGGAGGTGCAATGGCCATAGGCCATCCTTTGGGGGCTACA : 2597
PKT4 : TGTTTCAGGCATTTGCATCTCAGTTTGTTTATTGTCGGAACAAGCTCGGGCTAGACGGCGGAAAAGATCAATGTCAATGGAGGAGCCATAGCCATTGGACATCCCTTGGGGGCTACA : 2301
PKT2 : TAAAACAGGCATTTGCATCTCAGTATGTGTAATCTTGCAGAAAGTTAGAGCTGGATATGAAAAGGTCAATGTTAATGGAGGAGCCATTGCTATTGGCCATCCCTGGGTGCTACA : 2104
      <<12                                     12>

      *      2660          *      2680          *      2700          *      2720          *      2740          *      2760
PKT3 : GGTATATCTGTTAAGAACAGGACTAAAAGTGTAAAGCTTGTTCGAATTTACCAACCATAGAGACTTGAATTAATGAGAATCTCTCAACTCTCAACCAACCTTAAGAAATCCAC : 2712
PKT4 : GGTATGCTT----- : 2310
PKT2 : GGT----- : 2107
      >

```



```

      *      2780      *      2800      *      2820      *      2840      *      2860      *
PKT3 : AGCCGAGATATATCTTTATAAACTCTGTCTCAGTAATGGTTCTTTAACTCACTACAAGAGCTCTATCTTGCCGCAGTCCGTCCCTTTTCAAAGCATATCAATGTCTACATTGTG : 2827
PKT4 : -----CCTCTTGTGTCACATATCATTAAATT-----TCCTTCTCATCTAACAA----- : 2353
PKT2 : -----TACTAGCTAACGAAGCTGTCTAACTAATAAA----- : 2140

      2880      *      2900      *      2920      *      2940      *      2960      *      2980      *
PKT3 : TCTATATAATTATAAAACGCTCATGAACAGTTGCTAATACTTGTAAAACTCTCACAGGAGCGCGTGTGTGCTACATTGTTGCAAGGAGATGAAACGCCGTGGTAAAGACTGCC : 2942
PKT4 : -----CTACACATTCTCAGGAGCCAGATGCGTGTGCGACGCTGCTGCATGAGATGAAACGACGTGGTAAAGACTGTC : 2424
PKT2 : -----ATGCAACATTGTGTACAGCTAATAGAGTTCTTTCTATATGTTCAGGAGCTGGATGTGTGCGACATTGTTGCAAGGAGATGAAACGGAGAGGAAAAGATTGCC : 2244
                                <<13

      3000      *      3020      *      3040      *      3060      *      3080      *      3100
PKT3 : GTTTTGGAGTAGTGTCAATGTGCATTGGTATGCTTCTCACTCGAATAATTCAATGTGAATTTGGTGTCTGAAATAAGTGTCTGAGTGATGGTTTACTGATTGTGTAAACTCTGG : 3057
PKT4 : GTTTTGGCGTAGTGTCAATGTGTATAGGTATGTAATATGTTCT---ACTCTACCTCTCTTTGACTCCCACTCTCTCGTACACACTAAACACTCTCACTGTGACTCTGAAACCGCA : 2536
PKT2 : GCTTTGGAGTAATCTCAATGTGCATAGGTATATATAT-----AAGATTTTCTTTTTCAGATTGAATTGGATCAATGATTACCTATCTTAAATGGTATATGTAAATGCCAAGCTCA : 2352
                                13>>

      *      3120      *      3140      *      3160      *      3180      *      3200      *
PKT3 : CAGGGACGGGGATGGGTGCAGCAGCTGTGTTTGAGAGAGGAGATGCAGTTGATGAGCTTCGCAACGCAAGCAAAGTTGAACCGCTAGGTCTTTGTCCAAGGACGCTCG-- : 3167
PKT4 : CAGGTTCCGGGATGGGAGCAGCCGCTGTGTTTGAGAGAGGAGGCGGTGTGGATGAGCTCTCTGTATGTCCGAAAGT-----C-- : 2613
PKT2 : CAGGCCTGGGAATGGGAGCTGCAGCTGTTTTTGAGAGGGGAGACTCTGTGATAACTTGTCCAACGCTCTGTGGCTAACCGGGATAGTCAAT----- : 2444
                                <<14

```

Bibliography

- Abe, H., Ohtake, A., Yamamoto, S., Satoh, Y., Takayanagi, M., Amaya, Y., Takiguchi, M., Sakuraba, H., Suzuki, Y., Mori, M. & Niimi, H. (1993): Cloning and sequence analysis of a full length cDNA encoding human mitochondrial 3-oxoacyl-CoA thiolase. *Biochim. Biophys. Acta*, **1216**: 304-306.
- Abola, E.E., Sussman, J.L., Prilusky, J. & Manning, N.O. (1997). Protein Data Bank archives of three-dimensional macromolecular structures. In C. W. Carter Jr. & R. M. Sweet (Eds.), (pp. 556-571). San Diego: Academic Press.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990): Basic local alignment search tool. *J. Mol. Biol.*, **215**: 403-410.
- Anderson, M.L.M. & Young, B.D. (1985). Quantitative filter hybridization. In B. D. Hames & S. J. Higgins (Eds.), *Nucleic acid hybridization. A practical approach* (pp. 73-111). Oxford, Washington D.C.: IRL Press Ltd.
- André, D., Colau, D., Schell, J., van Montagu, M. & Hernalsteens, J.-P. (1986): Gene tagging in plants by a T-DNA insertion mutagen that generates APH (3')II-plant gene fusions. *Mol. Gen. Genet.*, **204**: 512-518.
- Anthony, R.G., James, P.E. & Jordan, P.R. (1993): Cloning and sequence analysis of a *flo/lfy* homologue isolated from *Brassica oleracea* L. var. *botrytis*. *Plant Mol. Biol.*, **22**: 1163-1166.
- Arakawa, H., Takiguchi, M., Amaya, Y., Nagata, S., Hayashi, H. & Mori, M. (1987): cDNA-derived amino acid sequence of rat mitochondrial 3-oxoacyl-CoA thiolase with no transient presequence: structural relationship with peroxisomal isozyme. *EMBO J.*, **6**: 1361-1366.
- Aristotle (1989). Generation of Animals. In J. Barnes (Eds.), *The complete works of Aristotle* Princeton: Princeton University Press.
- Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. & Struhl, K. (Eds.; 1996). *Current Protocols in Molecular Biology*. New York: Wiley Interscience.
- Azpiroz, R., Wu, Y., LoCascio, J.C. & Feldmann, K.A. (1998): An *Arabidopsis* brassinosteroid-dependent mutant is blocked in cell elongation. *Plant Cell*, **10**: 219-230.
- Bairoch, A. & Apweiler, R. (1998): The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucl. Acids Res.*, **26**: 38-42.

- Bairoch, A. & Bucher, P. (1994): PROSITE-Recent developments.
Nucl. Acids Res., **22**: 3583-3589.
- Baker, B., Schell, J., Lor, H. & Fedoroff, N. (1986): Transposition of the maize controlling element "Activator" in tobacco.
Proc. Natl. Acad. Sci. USA, **83**: 4844-4848.
- Bancroft, I., Bhatt, A.M., Sjodin, C., Scofield, S., Jones, J.D.G. & Dean, C. (1992): Development of an efficient two-element transposon tagging system in *Arabidopsis thaliana*.
Mol. Gen. Genet., **233**: 449-461.
- Barker, W.L., Garavelli, J.S., Haft, D.H., Hunt, L.T., Marzec, C.R., Orcutt, B.C., Srinivasarao, G.Y., Yeh, L.-S.L., Ledley, R.S., Mewes, H.-W., Pfeiffer, F. & Tsugita, A. (1998): The PIR-internet protein sequence database.
Nucl. Acids Res., **26**: 27-32.
- Barthels, N., Van der Lee, F., Klap, J., Goddijn, O.J.M., Ohl, S.A., Puzio, P., Grundler, F.M.W., Karimi, M., Robertson, L., Robertson, W.M., Lindsey, K., Van Montagu, M., Gheysen, G. & Sijmons, P. (1996): Nematode-inducible *A. thaliana* regulatory sequences driving reporter gene expression in nematode feeding structures.
Plant Cell, **9**: 2119-2134.
- Barton, M.K. & Poethig, R.S. (1993): Formation of the shoot apical meristem in *Arabidopsis thaliana*: and analysis of development in the wild type and in the *shoot meristemless* mutant.
Development, **119**: 823-831.
- Bell, C.J. & Ecker, J.R. (1994): Assignment of 30 microsatellite loci to the linkage map of *Arabidopsis*.
Genomics, **19**: 137-144.
- Belyavsky, A., Vinogradova, T. & Rajewsky, K. (1989): PCR-based cDNA library construction - General cDNA libraries at the level of a few cells.
Nucl. Acids Res., **17**: 2919-2932.
- Bennett, M.D. & Smith, J.B. (1976): Nuclear DNA amounts in angiosperms.
Phil. Trans. R. Soc. Lond., **B274**: 227-274.
- Benson, D.A., Boguski, M.S., Lipman, D.J. & Ostell (1997): GenBank.
Nucl. Acids Res., **25**: 1-6.
- Berninger, G., Schmidtchen, R., Casel, G., Knorr, A., Rautenstrauss, K., Kunau, W.H. & Schweizer, E. (1993): Structure and metabolic control of the *Yarrowia lipolytica* peroxisomal 3-oxoacyl-CoA-thiolase gene.
Eur. J. Biochem., **216**: 607-613.
- Bevan, M.W. (1984): Binary *Agrobacterium* vectors for plant transformation.
Nucl. Acids Res., **12**: 8711-8721.

- Bhatt, A.M., Page, T., Lawson, E.J.R., Lister, C. & Dean, C. (1996): Use of *Ac* as an insertional mutagen in *Arabidopsis*. *Plant J.*, **9**: 935-945.
- Bingham, P.M., Levis, R. & Rubin, G.M. (1981): Cloning of DNA sequences from the *white* locus of *D. melanogaster* by a novel and general method. *Cell*, **25**: 693-704.
- Bisgrove, S.R., Crouch, M.L. & Fernandez, D.E. (1995): Chimeric nature of precociously-germinating *Brassica napus* embryos: mRNA accumulation patterns. *J. Exp. Bot.*, **46**: 27-33.
- Blattner, F.R., Plunkett, G., F., I.M.G., Perna, N.T. & Glasner, F.D. (1997): Untitled. GenBank. Accession AE000460.
- Blumenthal, T. (1998): Gene clusters and polycistronic transcription in eukaryotes. *Bioessays*, **20**: 480-481.
- Bodnar, A.G. & Rachubinski, R.A. (1990): Cloning and sequence determination of cDNA encoding a second rat liver peroxisomal 3-ketoacyl-CoA thiolase. *Gene*, **91**: 193-199.
- Bojorquez, G. & Gomez-Lim, M.A. (1995): Peroxisomal thiolase mRNA is induced during mango fruit ripening. *Plant Mol. Biol.*, **28**: 811-820.
- Botstein, D., Skolnick, M. & Davis, R.W. (1980): Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.*, **32**: 314-331.
- Bout, A., Teunissen, Y., Hashimoto, T., Benne, R. & Tager J, M. (1988): Nucleotide sequence of human peroxisomal 3-oxoacyl-CoA thiolase. *Nucl. Acids Res.*, **16**: 10369.
- Breathnach, R. & Chambon, P. (1981): Organization and expression of eukaryotic split genes coding for proteins. *Ann. Rev. Biochem.*, **50**: 349-383.
- Breyne, P., De Loose, M., Dedonder, A., Van Montagu, M. & Depicker, A. (1993): Quantitative kinetic analysis of b-glucuronidase activities using a computer-directed microtiter plate reader. *Plant Mol. Biol. Rep.*, **11**: 21-31.
- Brown, C.M., Dalphin, M.E., Stockwell, P.A. & Tate, W.P. (1993): The translational termination signal database. *Nucl. Acids Res.*, **21**: 3119-3123.
- Brown, J.W., Smith, P. & Simpson, C.G. (1996): *Arabidopsis* consensus intron sequences. *Plant Mol. Biol.*, **32**: 531-535.

- Brownlee, C. & Berger, F. (1995): Extracellular matrix and pattern in plant embryos: on the lookout for developmental information. *Trends Genet.*, **11**: 344-348.
- Bruinenberg, P.G., Evers, M., Waterham, H.R., Kuipers, J., Arnberg, A.C. & Ab, G. (1989): Cloning and sequencing of the peroxisomal amine oxidase gene from *Hansenula polymorpha*. *Biochim. Biophys. Acta*, **1008**: 157-167.
- Bryan, E.M. (1995): Untitled. GenBank. Accession P45855.
- Burke, T.J., Callis, J. & Vierstra, R.D. (1988): Characterization of a polyubiquitin gene from *Arabidopsis thaliana*. *Mol. Gen. Genet.*, **213**: 435-443.
- Burr, F.A. & Burr, B. (1980). The cloning of zein sequences and an approach to zein genetics. In C. J. Leaver (Eds.), *Genome Organization and Expression in Plants* (pp. 227-231). Plenum Press.
- Busch, M., Mayer, U. & Jürgens, G. (1996): Molecular analysis of the *Arabidopsis* pattern formation of gene GNOM: gene structure and intragenic complementation. *Mol. Gen. Genet.*, **250**: 681-691.
- Busk, P.K. & Pages, M. (1998): Regulation of abscisic acid induced transcription. *Plant Mol. Biol.*, **37**: 425-435.
- Cecchini, E., J. D.P., Gery, C., Kaiser, K., SENTRY, J. & Milner, J.J. (1993): Identification of genes up-regulated in dedifferentiating *Nicotiana glauca* pith tissue, using an improved method for constructing a subtractive cDNA library. *Nucl. Acids. Res.*, **21**: 5742-5747.
- Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W. & Prasher, D.C. (1994): Green fluorescent protein as a marker for gene expression. *Science*, **263**: 802-805.
- Chardin, P. & Tavitrán, A. (1986): The *ral* gene: a new *ras* related gene isolated by the use of a synthetic probe. *EMBO J.*, **5**: 2203-2208.
- Chaubet, N., Flenet, M., Clement, B., Brignon, P. & Gigot, C. (1996): Identification of *cis* elements regulating the expression of an *Arabidopsis* histone H4 gene. *Plant J.*, **10**: 425-435.
- Choe, S., Dilkens, B.P., Fujioka, S., Takatsuto, S., Sakuriu, A. & Feldmann, F.A. (1998): The *DWF4* gene of *Arabidopsis* encodes a cytochrome P450 that mediates multiple 22 α -hydroxylation steps in brassinosteroid biosynthesis. *Plant Cell* **10**: 231-243

- Chory, J., Nagpal, P. & Peto, C. (1991): Phenotypic and genetic analysis of *det2*, a new mutant that affects light-regulated seedling development in *Arabidopsis*. *Plant Cell*, **3**: 445-459.
- Chuck, G., Robbins, T., Niggar, C., Ralston, E., courtney-Gutterson, N. & Dooner, H.K. (1993): Tagging and cloning of a petunia flower color gene with the maize transposable element *Activator*. *Plant Cell*, **5**: 371-378.
- Churchill, F.B. (1994). The rise of Classical Descriptive Embryology. In S. F. Gilbert (Eds.), *A conceptual history of modern embryology* (pp. 1-29). Baltimore and London: The John Hopkins University Press.
- Clarke, M.C., Wei, W. & Lindsey, K. (1992): High frequency transformation of *Arabidopsis thaliana* by *Agrobacterium tumefaciens*. *Plant Mol. Biol. Rep.*, **10**: 178-189.
- Clayton, C.E. (1985). Structure and regulated expression of genes encoding fructose bisphosphate aldolase in *Trypanosoma brucei*. *EMBO J.* **4**: 2997-3003.
- Cocke, R., Raynal, M., Laudie, M., Grellet, F., Delseny, M., Morris, P.L., Guerrier, D., Grandat, J., Quigley, F., Clabault, G. *et al* (1996): Further progress towards a catalogue of all *Arabidopsis* genes: analysis of a set of 5000 non-redundant ESTs. *Plant J.*, **9**: 101-124.
- Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry III, C.E., *et al.* (1998): Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**: 537-544.
- da Rocha, P.S.C.F. & Bertrand, H. (1995): Structure and comparative analysis of the rDNA intergenic spacer of *Brassica rapa*-Implications for the function and evolution of the *Cruciferae* spacer. *Eur. J. Biochem.*, **229**: 550-557.
- da Rocha, P.S.C.F., Topping, J.F. & Lindsey, K. (1995): Characterizing T-DNA tagged genes expressed in *Arabidopsis* embryos. *J. Exp. Bot.*, **46**: suppl. pp. 20.
- Dalphin, M.E., Brown, C.M., Stockwell, P.A. & Tate, W.P. (1998): The translational signal database, TransTerm, is now a relational database. *Nucl. Acids Res.*, **26**: 335-337.
- Dawe, R.K. & Freeling, M. (1991): Cell lineage and its consequences in higher plants. *Plant J.*, **1**: 3-8.
- de Hoop, M.J. & Geert, A.B. (1992): Import of proteins into peroxisomes and other microbodies. *Biochem J.*, **286**: 657-669.

de Jong, A.J., Cordewener, J., LoSchiavo, F., Terzi, M., Vandekerckhove, J., van Kammen, A. & de Vries, S. (1992): A carrot somatic embryo mutant is rescued by chitinase. *Plant Cell*, **4**: 425-433.

de Jong, A.J., Heidstra, R., Spaink, H.P., Hartog, M.V., Meijer, E.A., Hendriks, T., Lo Schiavo, F., Terzi, M., Bisseling, T., Van Kammen, A. & De Vries, S.C. (1993): *Rhizobium* lipooligosaccharides rescue a carrot somatic embryo mutant. *Plant Cell*, **5**: 615-620.

de Vet, E.C.J.M., van den Broek, B.T.E. & van den Bosch, H. (1997a). Nucleotide sequence of human alkylhydroxyacetonephosphate synthase cDNA reveals the presence of a peroxisomal targeting signal 2. *Biochim. Biophys. Acta* **1346**: 25-29.

de Vet, E.C.J.M., Lahaut, G.J.H.T.J. & van den Bosch, H. (1997b). Polymerase chain reaction-based cloning of alkyl-dihydroxyacetonephosphate synthase complementary DNA from guinea pig liver. *J. Biol. Chem.* **272**: 798-803.

Deleage, G. & Roux, B. (1987): An algorithm for protein secondary structure prediction based on class prediction. *Prot. Eng.*, **1**: 289-294.

Deleage, G. & Roux, B. (1987): The double prediction method - An algorithm for improving the accuracy of secondary structure prediction of proteins from class prediction. *Prot. Eng.*, **1**: 240.

Devereux, J., Haeberli, P. & Smithies, O. (1984): A comprehensive set of sequence-analysis programs for the Vax. *Nucl. Acids Res.*, **12**: 387-395.

Devic, M., Hecht, V., Berger, C., Lindsey, K., Delseny, M. & Gallois, P. (1995): Assessment of promoter trap as a tool to study zygotic embryogenesis in *Arabidopsis thaliana*. *C.R. Acad. Sci. Paris, Sciences de la vie*, **318**: 121-128.

Devos, K.M. & Gale, M.D. (1997): Comparative genetics in genomes. *Plant Mol. Biol.*, **35**: 3-15.

Dieuaide-Noubhani, M., Brouquisse, R., Pradet, A. & Raymond, P. (1992): Increased fatty acid β -oxidation after glucose starvation in maize root tips. *Plant Physiol.*, **99**: 595-600.

Dieuaide-Noubhani, M., Canioni, P. & Raymond, P. (1997): Sugar-starvation-induced changes of carbon metabolism in excised maize root tips. *Plant Physiol.*, **115**: 1505-1513.

DiRusso, C.C. (1990): Primary sequence of the *Escherichia coli* fadBA operon, encoding the fatty acid-oxidizing multienzyme complex, indicates a high degree of homology to eukaryotic enzymes. *J. Bacteriol.*, **172**: 6459-6468.

- Do, Y.-Y. & Huang, P.-L. (1997). Characterization of a pollination-related cDNA from *Phalaenopsis* encoding a protein which is homologous to human peroxisomal acyl-CoA oxidase. *Arch. Biochem. Biophys.* **344**: 295-300.
- Dodeman, V.L., Ducreux, G. & Kreis, M. (1997): Zygotic embryogenesis *versus* somatic embryogenesis. *Journal of Experimental Botany*, **48**: 1493-1509.
- Dolan, L., Janmaat, K., Willemsen, V., Linstead, P., Poethig, S., Roberts, K. & Scheres, B. (1993): Cellular organisation of the *Arabidopsis thaliana* root. *Development*, **119**: 71-84.
- Donald, R.G.K. & Cashmore, A.R. (1990): Mutation of either G box or I box sequences profoundly affects expression from the *Arabidopsis* rbcS-1A promoter. *EMBO J.*, **9**: 1717-1726.
- Donnison, I.S., Siroky, J., Vyskot, B., Saedler, H. & Grant, S.R. (1996): Isolation of Y-chromosome-specific sequences from *Silene latifolia* and mapping of male-sex determining genes using representational difference analysis. *Genetics*, **144**: 1893-1901.
- Doring, H.-P. & Starlinger, P. (1986): Molecular genetics of transposable elements in plants. *Ann. Rev. Genet.*, **20**: 175-200.
- Dower, W.J., Miller, J.F. & Ragsdale, C.W. (1988): High efficiency transformation of *E. coli* by high voltage electroporation. *Nucl. Acids Res.*, **16**: 6127-6145.
- Eccleston, V.S., Cranmer, A.M., Voelker, T.A. & Ohlrogge, J.B. (1996): Medium-chain fatty acid biosynthesis and utilization in *Brassica napus* plants expressing lauroyl-acyl carrier protein. *Planta*, **198**: 46-53.
- Eccleston V.S. et al (1998): Expression of lauroyl-acyl carrier protein thioesterase in *Brassica napus* seeds induces pathways for both fatty acid oxidation and biosynthesis and implies a set point for triacylglycerol accumulation. *Plant Cell* **10**: 613-621.
- Edwards, J.B.D.M., Delort, J. & Mallet, J. (1991): Oligodeoxyribonucleotide ligation to single-stranded cDNAs - new tool for cloning 5'-ends of messenger-RNAs and for constructing cDNA libraries by in vitro amplification. *Nucl. Acids Res.*, **19**: 5227-5232.
- Evan, G.I., Lewis, G.K., Ramsay, G. & Bishop, J.M. (1985): Isolation of monoclonal antibodies specific for the human c-myc proto-oncogene product. *Mol. Cell. Biol.*, **5**: 3610-3616.
- Favery, B., Lecomte, P., Bechtold, N., Boudez, D., Dalmaso, A. & Asad, P. (1998): RPE a plant gene involved in early developmental steps of nematode feeding cells. *EMBO J.*, **17**: 6799-6811.

- Fedoroff, N.V., Furtak, D.B. & Nelson, G.E., Jr. (1984): Cloning of the *bronze* locus in maize by a simple and generalizable procedure using the transposable controlling element *Activator*. *Proc. Natl. Acad. Sci. USA*, **81**: 3825-3829.
- Feinberg, A.P. & Vogelstein, B. (1984): A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.*, **137**: 266-267.
- Feldmann, K.A. (1991): T-DNA insertion mutagenesis in *Arabidopsis*: mutational spectrum. *Plant J.*, **1**: 71-82.
- Feldmann, K.A. & Marks, M.D. (1987): *Agrobacterium*-mediated transformation of germinating seeds of *Arabidopsis thaliana*: a non-tissue culture approach. *Mol. Gen. Genet.*, **208**: 1-9.
- Fernandez, D.E., Turner, F.R. & Crouch, M.L. (1991): *In situ* localization of storage protein mRNAs in developing meristems of *Brassica napus* embryos. *Development*, **111**: 299-313.
- Ferreira da Rocha, P.S.C., Topping, J.F. & Lindsey, K. (1996): Promoter trapping in *Arabidopsis*: One T-DNA tag, three transcripts and two thiolase genes. *J. Exp. Bot.*, **47**: suppl. pp. 24.
- Ferreira, P.C.G., Hemerly, A.S., Villaroel, R., Montagu, M.V. & Inze, D. (1991): The *Arabidopsis* functional homolog of the p34^{cdc2} protein kinase. *Plant Cell*, **3**: 531-540.
- Fields, S. & Song, O.-K. (1989): A novel genetic system to detect protein-protein interactions. *Nature*, **340**: 245-246.
- Flanders, D.J., Weng, S., Petel, F.X. & Cherry, J.M. (1998): AtDB, the *Arabidopsis thaliana* database, and graphical-web-display of progress by the *Arabidopsis* genome initiative. *Nucl. Acids Res.*, **26**: 80-84.
- Flavell, R. (1980): The molecular characteristics and organization of plant chromosomal sequences. *Ann. Rev. Plant Physiol.*, **31**: 569-596.
- Flynn, C.R., Mullen, R.T. & Trelease, R.N. (1998): Mutational analysis of a type 2 peroxisomal targeting signal that is capable of directing oligomeric protein import into tobacco BY-2 gyoxisomes. *Plant J.*, **16**: 709-720.
- Fobert, P.R., Labbé, H., Cosmopoulos, J., Gottlob-McHugh, S., Ouellet, T., Hattori, J., Sunohara, G., Iyer, V.N. & Miki, B.L. (1994): T-DNA tagging of a seed coat-specific cryptic promoter in tobacco. *Plant J.*, **6**: 567-577.

- Fobert, P.R., Miki, B.L. & Iyer, V.N. (1991): Detection of gene regulatory signals in plants revealed by T-DNA-mediated fusions.
Plant Mol. Biol., **17**: 837-851.
- Forsthoefel, N.R., Wu, Y., Schulz, B., Bennett, M.J. & Feldmann, K.A. (1992): T-DNA insertion mutagenesis in *Arabidopsis*: prospects and perspectives.
Aust. J. Plant Physiol., **19**: 353-366.
- Franzmann, L.H., Yoon, E.S. & Meinke, D.W. (1995): Saturating the genetic map of *Arabidopsis thaliana* with embryonic mutations.
Plant J., **7**: 341-350.
- Frishman, D. & Argos, P. (1996): Incorporation of non-local interactions in protein secondary structure prediction from the amino acid prediction.
Prot. Eng., **9**: 133-142.
- Fritze, K., Czaja, I. & Walden, R. (1995): T-DNA tagging of genes influencing polyamine metabolism: isolation of mutant plant lines and rescue of DNA promoting growth in the presence of a polyamine biosynthetic inhibitor.
Plant J., **7**: 261-271.
- Frohman, M.A., Dush, M.K. & Martin, G.R. (1988): Rapid production of full-length cDNAs from rare transcripts -amplification using a single gene-specific oligonucleotide primer.
Proc. Natl. Acad. Sci. USA, **85**: 8998-9002.
- Fukao, T., Kamijo, K., Osumi, T., Fujiki, Y., Yamaguchi, S., Orii, T. & Hashimoto, T. (1989): Molecular cloning and nucleotide sequence of cDNA encoding the entire precursor of rat mitochondrial acetoacetyl-CoA thiolase.
J. Biochem., **106**: 197-204.
- Fukao, T., Yamaguchi, S., Kano, M., Orii, T., Fujiki, Y., Osumi, T. & Hashimoto, T. (1990): Molecular cloning and sequence of the complementary DNA encoding human mitochondrial acetoacetyl-coenzyme A thiolase and study of the variant enzymes in cultured fibroblasts from patients with 3-ketothiolase deficiency.
J. Clin. Invest., **86**: 2086-2092.
- Garnier, J., Gibrat, J.-F. & Robson, B. (1996). GOR secondary structure prediction method version IV.
Meth. Enz., **266**: 540-553.
- Gasser, C.S., Budelier, K.A., Smith, A.G., Shah, D.M. & Fraley, R.T. (1989): Isolation of tissue-specific cDNAs from tomato pistils.
Plant Cell, **1**: 15-24.
- Gebhardt, J.S., Wadsworth, G.J. & Mathews, B. (1998). Characterization of a single soybean cDNA encoding cytosolic and glyoxysomal isozymes of aspartate aminotransferase.
Plant Mol. Biol. **37**: 99-108.

- Geourjon, C. & Deleage, G. (1994): SOPM : A self-optimised prediction method for protein secondary structure.
Prot. Eng., **7**: 157-164.
- Geourjon, C. & Deleage, G. (1995): SOPMA : Significant improvements in protein secondary structure prediction by prediction from multiple alignments.
Comput. Applic. Biosci., **11**: 681-684.
- Gerhardt, B. (1983): Localization of b-oxidation enzymes isolated from non-fatty plant tissues.
Planta, **159**: 238-246.
- Gheysen, G., Villaroel, R. & Van Montagu, M. (1991): Illegitimate recombination in plants: a role for T-DNA integration.
Genes Devel., **5**: 287-297.
- Ghosh, D. (1991): New developments of a transcription factors database.
Trends in Biochemical Sciences, **16**: 445-447.
- Gibrat, J.F., Garnier, J. & Robson, B. (1987): Further developments of protein secondary structure prediction using information theory - New parameters and consideration of residue pairs.
J. Mol. Biol., **198**: 425-443.
- Gietl, C. (1990): Glyoxysomal malate-dehydrogenase from watermelon is synthesized with an amino-terminal transit peptide.
Proc. Natl. Acad. Sci. USA, **87**: 5773-5777.
- Gietl, C. (1996): Protein targeting and import into plant peroxisomes.
Physiol. Plant., **97**: 599-608.
- Gilbert, S.F. (1991). *Developmental Biology* (3rd. Edition ed.). Sunderland: Sinauer Associates, Inc.
- Gilbert, S.F. (1994a). Induction and the origins of developmental genetics. In S. F. Gilbert (Eds.), *A conceptual history of modern embryology* (pp. 181-206). Baltimore and London: The John Hopkins University Press.
- Gilbert, S.F. (1994b). The origins of *Entwicklungsmechanik*. In S. F. Gilbert (Eds.), *A conceptual history of modern embryology* (pp. 43-61). Baltimore and London: The John Hopkins University Press.
- Giraudat, J., Hauge, B., Valon, C., Smalle, J., Parcy, F. & Goodman, H. (1992): Isolation of the Arabidopsis *ABI3* gene by positional cloning.
Plant Cell, **4**: 1251-1261.
- Glover, J., Andrews, D.W. & Rachubinski, R.A. (1994): Saccharomyces-cerevisiae peroxisomal thiolase is imported as a dimer.
Proc. Natl. Acad. Sci. USA, **91**: 10541-10545.

- Goldberg, R.B., Barker, S.J. & Perez-Grau, L. (1989): Regulation of gene expression during plant embryogenesis.
Cell, **56**: 149-160.
- Goodwin, P.B. (1978). Phytohormones and the development of higher plants. In D. S. Letham, P. B. Goodwin & T. J. V. Higgins (Eds.), *Phytohormones and related compounds-a comprehensive treatise* (pp. 31-173). Amsterdam, Oxford, New York: Elsevier/North Holland Biomedical Press.
- Gossler, A., Joyner, A.L., Rossant, J. & Skarnes, W.C. (1989): Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes.
Science, **244**: 463-465.
- Graham, I., Leaver, C.J. & Smith, S.M. (1992): Induction of malate synthase gene expression in senescent and detached organs of cucumber.
Plant Cell, **4**: 349-357.
- Grevelding, C., Becker, D., Kunze, R., Von Menges, A., Fantes, V., Schell, J. & Masterson, R. (1992): High rates of *Ac/Ds* germinal transposition in *Arabidopsis* suitable for gene isolation by insertional mutagenesis.
Proc. Natl. Acad. Sci. USA, **89**: 6085-6089.
- Grevelding, C., Fantes, V., Kemper, E., Schell, J. & Materson, R. (1993): Single-copy T-DNA insertions in *Arabidopsis* are the predominant form of integration in root-derived transgenics, whereas multiple insertions are found in leaf discs.
Plant Mol. Biol., **23**: 847-860.
- Gribskov, M. & Burgess, R.R. (1986): Sigma factors from *E. coli*, *B. subtilis*, phage SP01, and phage T4 are homologous.
Nucl. Acids Res., **14**: 6745-6763.
- Grierson, C., Du, J.-S., Zabala, M.T., Beggs, K., Smith, C., Holdsworth, M. & Bevan, M. (1994): Separate *cis* sequences and *trans* factors direct metabolic and developmental regulation of a potato tuber storage protein gene.
Plant J., **5**: 815-826.
- Genetics Computer Group (1994). *Program Manual for the Wisconsin Package, Version 8*. Madison, Wisconsin: Genetics Computer Group, Wisconsin.
- Guex, N., Henry, H., Flach, J., Richter, H. & Widmer, F. (1995): Glyoxysomal malate dehydrogenase and malate synthase from soybean cotyledons (*Glycine max* L.): enzyme association, antibody production and cDNA cloning.
Planta, **197**: 369-375.
- Guex, N. & Peitsch, M.C. (1997): The Swiss-PdbViewer and SWISS-MODEL: an environment for comparative protein modeling.
Electrophoresis, **18**: 2714-2723.

- Guiltinan, M.J., Marcotte, W.R. & Quatrano, R.S. (1990): A plant leucine zipper protein that recognizes an abscisic acid response element. *Science*, **250**: 267-270.
- Guyer, M.S. (1978): The gamma delta sequence of F is an insertion sequence. *J. Mol. Biol.*, **126**: 347-365.
- Haizel, F., Merckle, T., Pay, A., Fejes, E. & Naji, F. (1997): Characteristics of proteins that interact with the GTP-bound form of the regulatory GTPase Ran in *Arabidopsis*. *Plant J.*, **11**: 93-103.
- Hamburgh, M. (1971). *Theories of Differentiation*. London: Edward Arnold.
- Hardtke, C.S. & Berleth, T. (1998): The *Arabidopsis* gene *MONOPTEROS* encodes a transcription factor encoding embryo axis formation and vascular development. *EMBO J.*, **17**: 1405-1411.
- Harwood, C.S., Nichols, N.N., Kim, M.K., Ditty, J.L. & Parales, R.E. (1994): Identification of the *pcaRKF* gene cluster from *Pseudomonas putida*: involvement in chemotaxis, biodegradation, and transport of 4-hydroxybenzoate. *J. Bacteriol.*, **176**: 6479-6488.
- Hayashi, H., Bellis, L.D., Yamaguchi, K., Kato, A., Hayashi, M. & Nishimura, M. (1998). Molecular characterization of a glyoxysomal acyl-CoA oxidase that is synthesized as a precursor of higher molecular mass in pumpkin. *J. Biol. Chem.* **273**: 8301-8307.
- Hayashi, M., Toriyama, T., Kondo, M. & Nishimura, M. (1998): 2,4-Dichlorophenoxybutyric acid-resistant mutants of *Arabidopsis* have defects in glyoxysomal fatty acid β -oxidation. *Plant Cell*, **10**: 183-195.
- Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J. & Rouze, P. (1996): Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucl. Acids Res.*, **24**: 3439-3452.
- Heck, G.R., Nichols, K.W. & Fernandez, D.E. (1995): AGL15, a MADS domain protein expressed in developing embryos. *Plant Cell*, **7**: 1271-1282.
- Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A.E., Kel, O.V., Ignatieva, E.V., Ananko, E.A., Podkolodnaya, O.A., Kolpakov, F.A., Podkolodny, N.L. & Kolchanov, N.A. (1998): Databases of transcriptional regulation: TRANSFAC, TRRD, and COMPEL. *Nucl. Acids Res.*, **26**: 362-367.
- Henikoff, S. & Henikoff, J.G. (1992): Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**: 10915-10919.
- Herman, L., Jacobs, A., Van Montagu, M. & Depicker, A. (1990): Plant chromosome/marker gene fusion assay for study of normal and truncated T-DNA integration events. *Mol. Gen. Genet.*, **224**: 248-256.

- Higo, K., Ugawa, Y., Iwamoto, M. & Higo, H. (1998): PLACE: a database of plant *cis*-acting regulatory elements. Nucl. Acids Res., **26**: 358-359.
- Hijikata, M., Ishii, N., Kagamiyama, H., Osumi, T. & Hashimoto, T. (1987): Structural analysis of cDNA for rat peroxisomal 3-ketoacyl-CoA thiolase. J. Biol. Chem., **262**: 8151-8158.
- Hijikata, M., Wen, J.K., Osumi, T. & Hashimoto, T. (1990): Rat peroxisomal 3-ketoacyl-CoA thiolase gene. Occurrence of two closely related but differentially regulated genes. J. Biol. Chem., **265**: 4600-4606.
- Hiser, L., Basson, M.E. & Rine, J. (1994): ERG10 from *Saccharomyces cerevisiae* encodes acetoacetyl-CoA thiolase. J. Biol. Chem., **269**: 31383-31389.
- Hobbs, S.L.A., De Kpodar, P. & Long, C.M. (1990): The effect of T-DNA copy number, position and methylation on reporter gene expression in tobacco transformants. Plant Mol. Biol., **15**: 851-864.
- Hooks, M.A., Fleming, Y., Carson, T.R., Graham, I.A. (1999). No induction of β -oxidation in leaves of *Arabidopsis* that overproduce lauric acid. Planta **207**: 385-392.
- Hwang, Y., Karrer, E., BR, T., Chen, L. & Rodriguez, C. (1998): Three *cis*-elements required for rice alpha-amylase Amy3D expression during sugar starvation. Plant Mol. Biol., **36**: 331-341.
- Igual, J.C., Gonzalez-Bosch, C., DoPazo, J. & Perez-Ortin, J.E. (1992): Phylogenetic analysis of the thiolase family-implications for the evolutionary origin of peroxisomes. J. Mol. Evol., **35**: 147-155.
- Igual, J.C., Matallana, E., Gonzalez-Bosch, C., Franco, L. & Perez-Ortin, J.E. (1991): A new glucose-repressible gene identified from the analysis of chromatin structure in deletion mutants of yeast *SUC2* locus. Yeast, **7**: 379-389.
- Itoh, T., Aiba, H., Buba, T., Hayashi, K., Inada, T., Isono, K., Kasai, H., Kimura, S., Kitakawa, M., Kitagawa, M. & al., e. (1996): A 460-kb DNA sequence of the *escherichia coli* K-12 genome corresponding to the 40.1-50.0 min region on the linkage map. DNA Res., **3**: 441-445.
- Jefferson, R.A., Kavanagh, T.A. & Bevan, M.W. (1987): GUS fusions: b-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. EMBO J., **6**: 3901-3907.
- Jensen, E.U., Paludan, K., Hylgid-Nielsen, J.J., Jorgensen, P. & Marcker, K.A. (1981): The structure of a chromosomal leghaemoglobin gene from soybean. Nature, **291**: 677-679.

- Jiang, C., Iu, B. & Singh, J. (1996): Requirement of a CCGAC cis-acting element for cold induction of BN115 from winter *Brassica napus*. *Plant Mol. Biol.*, **30**: 679-684.
- Jorgensen, R., Snyder, C. & Jones, J.D.G. (1987): T-DNA is organized predominantly in inverted repeat structures in plants transformed with *Agrobacterium tumefaciens* C58 derivatives. *Mol. Gen. Genet.*, **207**: 471-477.
- Jürgens, G., Mayer, U., Busch, M., Lukowitz, W. & Laux, T. (1995): Pattern formation in the *Arabidopsis* embryo: a genetic perspective. *Phil. Trans. R. Soc. Lond. B*, **350**: 19-25.
- Jürgens, G., Mayer, U., Torres-Ruiz, R.A. & Bertleth, T. (1991): Genetic analysis of pattern formation in the *Arabidopsis* embryo. *Development*, **91** (Suppl. 1): 27-38.
- Kamijo, T., Aoyama, T., Komiyama, A. & Hashimoto, T. (1994): Structural analysis of cDNAs for subunits of human mitochondrial fatty acid beta-oxidation trifunctional protein. *Biochem. Biophys. Res. Commun.*, **199**: 818-825.
- Kamijo, T., Aoyama, T., Miyazaki, J. & Hashimoto, T. (1993): Molecular cloning of the cDNAs for the subunits of rat mitochondrial fatty acid beta-oxidation multienzyme complex. Structural and functional relationships to other mitochondrial and peroxisomal beta-oxidation enzymes. *J. Biol. Chem.*, **268**: 26452-26460.
- Kaminaka, H. & Tanaka, K. (1996): *Oryza sativa* glyoxysomal Malate Dehydrogenase. GenBank. Accession D85783.
- Kanayama, N., Ueda, M., Atomi, H., Kurihara, T., Kondo, J., Teranishi, Y. & Tanaka, A. (1994): Comparison of molecular structures and regulation of biosynthesis of unique thiolase isozymes localized only in peroxisomes of n-alkane-utilizing yeast, *Candida tropicalis*. *J. Ferment. Bioeng.*, **78**: 273-278.
- Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M. & Tabata, S. (1996): Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.*, **3**: 109-136.
- Katinka, M.D. (1997): Untitled. GenBank. Accession AJ003066.
- Kato, A., Hayashi, M., Kondom M. & Nishimura, M. (1996b): Targeting and processing of a chimeric protein with the N-terminal presequence of the precursor to glyoxysomal citrate synthase. *Plant Cell* **8**: 1601-1611

- Kato, A., Hayashi, M., Mori, H. & Nishimura, M. (1995): Molecular characterization of a glyoxysomal citrate synthase that is synthesized as a precursor of higher molecular mass in pumpkin.
Plant Mol. Biol. **27**: 377-390.
- Kato, A., Hayashi, M., Takeuchi, Y. & Nishimura, M. (1996): cDNA cloning and expression of a gene for 3-ketoacyl-CoA thiolase in pumpkin cotyledons.
Plant Mol. Biol., **31**: 843-852.
- Kato, A., Takeda-Yoshikawa, Y., Hayashi, M., Kondo, M., Hara-Nishimura, I. & Nishimura, M. (1998): Glyoxysomal malate dehydrogenase in pumpkin: Cloning of a cDNA and functional analysis of its presequence.
Plant Cell Physiol., **39**: 186-195.
- Kauschmann, A., Jessop, A., Koncz, C., Szekeres, M., Willmitzer, L. & Altmann, T. (1996): Genetic evidence for an essential role of brassinosteroids in plant development.
Plant J., **9**: 701-713.
- Kertbundit, S., De Greve, H., DeBoeck, F., Van Montagu, M. & Hernalsteens, J.-P. (1991): *In vivo* random b-glucuronidase gene fusions in *Arabidopsis thaliana*.
Proc. Natl. Acad. Sci. USA, **88**: 5212-5216.
- Kim, D.J. & Smith, S.M. (1994): Expression of a single gene encoding microbody NAD-malate dehydrogenase during glyoxysome and peroxisome development in cucumber.
Plant Mol. Biol., **26**: 1833-1841.
- Kim, S.Y., Chung, H.-J. & Thomas, T.L. (1997): Isolation of a novel class of bZIP transcription factors that interact with ABA-responsive and embryo-specification elements in the DC3 promoter using a modified yeast one-hybrid system.
Plant J., **11**: 1237-1251.
- Kimmel, A.R. & Berger, S.L. (1987). Preparation of cDNA and the generation of cDNA libraires: overview. In *Guide to Molecular Cloning Techniques* (pp. 307-316).
- Kitamura, T. (1996): Untitled.
GenBank. Accession X95443.
- Klahre, U., Noguchi, T., Fujioka, S., Takatsuto, S., Yokota, T., Nomura, T., Yoshida, S. & Chua, N.-H. (1998): The *Arabidopsis* *DIMINUTO/DWARF1* gene encodes a protein involved in steroid synthesis.
Plant Cell, **10**: 1677-1690.
- Klenk, H.P., Clayton, R.A., Tomb, J., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D. *et al.* (1997): The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*.
Nature, **390**: 364-370.
- Koltunow, A.M., Bicknell, R.A. & Chaudhury, A.M. (1995): Apomixis: molecular strategies for the generation of genetically identical seeds without fertilization.
Plant Phys., **108**: 1345-1352.

- Koncz, C., Martini, N., Mayerhofer, R., Koncz-Kalman, Z., Körber, H., Rédei, G.P. & Schell, J. (1989): High frequency T-DNA-mediated gene tagging in plants.
Proc. Natl. Acad. Sci. USA, **86**: 8467-8471.
- Konieczny, A. & Ausubel, F.M. (1993): A procedure to map *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers.
Plant J., **4**: 403-410.
- Koornneef, M., Jorna, M.L., Brinkhorst-van der Swan, D.C.C. & Karssen, C.M. (1982): The isolation of abscisic acid (ABA) deficient mutants by selection of induced revertants in non-germinating gibberellin sensitive lines of *Arabidopsis thaliana* (L) Heynh.
Theor. Appl. Genet., **61**: 385-393.
- Kowalchuk, G.A., Hartnett, G.B., Benson, A., Houghton, J.E., Ngai, K.L. & Ornston, L.N. (1994): Contrasting patterns of evolutionary divergence within the *Acinetobacter calcoaceticus* *pca* operon.
Gene, **146**: 23-30.
- Kowalski, S.P., Lan, T.-H., Feldmann, K.A. & Paterson, A.H. (1994): Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals islands of conserved organization.
Genetics, **138**: 499-510.
- Kurihara, T., Ueda, M., Kanayama, M., Kondo, J., Teranishi, Y. & Tanaka, A. (1992): Peroxisomal acetoacetyl-CoA thiolase of an N-alkane-utilizing yeast, *Candida tropicalis*.
Eur. J. Biochem., **210**: 999-1005.
- Lange, T., Hedden, P. & Graebe, J.E. (1994): Expression cloning of a gibberellin 20-oxidase, a multifunctional oxidase involved in gibberellin biosynthesis.
Proc. Natl. Acad. Sci. USA, **91**: 8552-8556.
- Laskowski, R.A., Hutchinson, E.G., Michie, A.D., Wallace, A.C., Jones, M.L. & Thornton, J.L. (1997): A Web-based database of summaries and analyses of all PDB structures.
Trends Biochem. Sci., **22**: 488-490.
- Lazarow, P.B. (1978): Rat liver peroxisomes catalyze the β oxidation of fatty acids.
J. Biol. Chem., **253**: 1522-1528.
- Learned, R.M. & Fink, G.R. (1989): 3-hydroxy-3-methylglutaryl-coenzyme A reductase from *Arabidopsis thaliana* is structurally distinct from the yeast and animal enzymes.
Proc. Natl. Acad. Sci. USA, **86**: 2779-2783.
- Lee, I., Rhee, Y.H. & Kim, J.Y. (1997): Cloning and sequencing of *phaA* from *Alcaligenes* sp SH-69.
GenBank. Accession AF002013.

- Lessard, P.A., Allen, R.D., Bernier, F., Crispino, J.D., Fujiwara, T. & Beachy, R.N. (1991): Multiple nuclear factors interact with upstream sequences of differentially regulated beta-conglycinin genes. *Plant Mol. Biol.*, **16**: 397-413.
- Levin, J.M., Robson, B. & Garnier, J. (1986): An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.*, **205**: 303-308.
- Li, J., Biswas, M.G., Chao, A., Russell, D.W. & Chory, J. (1997): Conservation of function between mammalian and plant 5a-reductase. *Proc. Natl. Acad. Sci. USA*, **94**: 3554-3559.
- Liang, P. & Pardee, A.B. (1992): Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, **257**: 967-971.
- Liang, X., Abel, S., Keller, J.A., Shen, N.F. & Theologis, A. (1991): The 1-aminocyclopropane-1-carboxylase synthase gene family of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA*, **89**: 11046-11050.
- Liebergesell, M. & Steinbuchel, A. (1993): Cloning and molecular analysis of the poly(3-hydroxybutyric acid) biosynthetic genes of *Thiocystis violacea*. *Appl. Microbiol. Biotechnol.*, **38**: 493-501.
- Lindsey, K. & Topping, J.F. (1993): Embryogenesis - a question of pattern. *J. Exp. Bot.*, **44**: 359-374.
- Lindsey, K., Topping, J.F., da Rocha, P.S.C.F., Horne, K.L., Muskett, P.R., May, V.J. & Wei, W. (1996). Insertional mutagenesis to dissect embryonic development in *Arabidopsis*. In T. Wang & A. C. Cuming (Eds.), *Embryogenesis - Generation of a Plant* (pp. 51-76). Oxford: Bios Scientific Publishers.
- Lindsey, K., Wei, W., Clarke, M.C., McArdle, H.F., Rooke, L.M. & Topping, J.F. (1993): Tagging genomic sequences that direct transgene expression by activation of a promoter trap in plants. *Trans. Res.*, **2**: 33-47.
- Lisitnyn, N., Lisitnyn, N. & Wegler, M. (1993): Cloning differences between two complex genomes. *Science*, **259**: 946-951.
- Lister, C. & Dean, C. (1993): Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. *Plant J.*, **4**: 745-750.
- Liu, C.-M., Xu, Z.-H. & Chua, N.-C. (1993): Proembryo culture: *in vitro* development of early globular-stage zygotic embryos from *Brassica juncea*. *Plant J.*, **3**: 291-300.

- Liu, Y.-G., Mitsukawa, N., Vazquez-Tello, A. & Witthier, R.F. (1995): Generation of a high-quality P1 library of *Arabidopsis* suitable for chromosome walking. *Plant J.*, **7**: 351-358.
- Logemann, E., Parniske, M. & Hahlbrock, K. (1995): Modes of expression and common structural features of the complete phenylalanine ammonia-lyase gene family in parsley. *Proc. Natl. Acad. Sci. USA*, **92**: 5905-5909.
- Long, D., Goodrich, J., Wilson, K., Sundberg, E., Martin, M., Puangsomlee, P. & Coupland, G. (1997): *Ds* elements on all five *Arabidopsis* chromosomes and assessment of their utility for transposon tagging. *Plant J.*, **11**: 145-148.
- Long, D., Swinborne, J., Martin, M., Wilson, K., Sundberg, E., Lee, K. & Coupland, G. (1993): Analysis of the frequency of inheritance of transposed *Ds* elements in *Arabidopsis* after activation by a CaMV 35S promoter fusion to the *Ac* transposase gene. *Mol. Gen. Genet.*, **241**: 627-636.
- Long, J.A., Moan, E.I., Medford, J.I. & Barton, M.K. (1996): A member of the KNOTTED class of homeodomain proteins encoded by the *STM* gene of *Arabidopsis*. *Nature*, **379**: 66-69.
- Lu, G., Delisle, A.J., van Wiemeersch, L., Goodman, H.M. & van Montagu, M. (1992): Brain proteins in plants: An *Arabidopsis* homolog to neurotransmitter pathway activators is part of a DNA binding complex. *Proc. Natl. Acad. Sci. USA*, **89**: 11490-11494.
- Lu, P., Porat, R., Nadeau, J.A. & O'Neill, S.D. (1996): Identification of a meristem L1 layer-specific gene in *Arabidopsis* that is expressed during embryonic pattern formation and defines a new class of homeobox genes. *Plant Cell*, **8**: 2155-2168.
- Lumbreras, V. (1995): The use of alternative promoter in the *Arabidopsis thaliana* *HMG1* gene generates an mRNA that encodes a novel 3-hydroxy-3-methylglutaryl coenzyme A reductase isoform with an extended N-terminal region. *Plant J.*, **8**: 541-549.
- Makoto, H., Ishii, N., Kagamiyama, H., Osumi, T. & F, T.H.F. (1987): Structural analysis of cDNA for rat peroxisomal 3-ketoacyl-CoA thiolase. *J. Biol. Chem.*, **262**: 8151-8148.
- Mansfield, S.G. & Briarty, L.G. (1991): Early embryogenesis in *Arabidopsis thaliana*. II. The developing embryo. *Can. J. Bot.*, **69**: 461-476.
- Marks, M.D. & K.A. Feldmann (1989): Trichome development in *Arabidopsis thaliana*. I. T-DNA tagging of the *GLABROUS 1* gene. *Plant Cell*, **1**: 1043-1050.

- Masamune, S., Palmer, M.A., Gamboni, R., Thompson, S., Davies, J.T., Williams, S.F., Peoples, O.P., Sinskey, A.J. & Walsh, C.T. (1989): Bio-Claissen condensation catalyzed by thiolase from *Zoogloea ramigera*. Active site cysteine residues. *J. Am. Chem. Soc.*, **111**: 1879-1881.
- Mathieu, M., Modis, Y., Zeelen, J.P., Engel, C.K., Abagyan, R.A., Ahlberg, A., Rasmussen, B., Lamzin, V.S., Kunau, W.H. & Wierenga, R.K. (1997): The 1.8 angstrom crystal structure of the dimeric peroxisomal thiolase of *saccharomyces cerevisiae*: implications for substrate binding and reaction mechanism. *J. Mol. Biol.*, **273**: 714-728.
- Mathieu, M., Zeelen, J.P., Pauptit, R.A., Erdmann, R., Kunau, W.H. & Wierenga, R.K. (1994): The 2.8-angstrom crystal-structure of peroxisomal 3-ketoacyl-coa thiolase of *Saccharomyces cerevisiae*-a 5 layered alpha-beta-alpha-beta-alpha-structure constructed from 2 core domains of identical topology. *Structure*, **2**: 797-808.
- Mayer, U., Büttner, G. & Jürgens, G. (1993): Apical-basal pattern formation in the *Arabidopsis* embryo: studies on the role of the *gnom* gene. *Development*, **117**: 149-162.
- Mayer, U., Torres-Ruiz, R.A., Berleth, T., Miséra, S. & Jürgens, G. (1991): Mutations affecting body organization in the *Arabidopsis* embryo. *Nature*, **353**: 402-407.
- McCarty, D.R. & Carson, C.B. (1991): The molecular genetics of seed maturation in maize. *Physiol. Plant.*, **81**: 267-272.
- Meinke, D.W. (1985): Embryo-lethal mutants of *Arabidopsis thaliana*: analysis of mutants with a wide range of lethal phases. *Theor. Appl. Genet.*, **69**: 543-552.
- Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D. & Koornneef, M. (1998): *Arabidopsis thaliana*: A model plant for genome analysis. *Science*, **282**: 662-682.
- Menkens, A.E., Schindler, U. & Cashmore, A.R. (1995): The G box: a ubiquitous regulatory DNA element in plants bound by the GBF family of bZip proteins. *Trends Biochem Sci.*, **20**: 506-510.
- Meyerowitz, E.M. (1987): *Arabidopsis thaliana*. *Ann. Rev. Genet.*, **21**: 93-111.
- Meyerowitz, E.M. (1992). Introduction to the *Arabidopsis* genome. In C. Koncz, N.-M. Chua & J. Schell (Eds.), *Methods in Arabidopsis research* (pp. 100-119). World Scientific Publishing.
- Michelmore, R.W., Paran, I. & Kesseli, R.V. (1991): Identification of markers linked to disease resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. USA*, **88**: 9828-9832.

- Middleton, B. (1973): The oxoacyl-Coenzyme A of animal tissues. *Biochem J.*, **132**: 717-730.
- Mihalik, S.J., Morrell, J.C., Kim, D., Sachsteder, K.A., Watkins, P.A. & Gould, J. (1997). Identification of PAHX, a Refsum disease gene. *Nature Genet.* **17**: 185-189.
- Mikami, K., Nakayama, T., Kawata, T., Tabata, T. & Iwabuchi, M. (1989): Specific interaction of nuclear protein HBP-1 with the conserved hexameric sequence ACGTCA in the regulatory region of wheat histone genes. *Plant Cell Physiology*, **30**: 107-119.
- Miller, S.S., Driscoll, B.T., Gregerson, R.G., Vance C.P. & Gant, J.s. (1997): Alfalfa Malate Dehydrogenase (MDH): Molecular cloning and characterization of five different forms, including a unique nodule-enhanced MDH. GenBank. Accession AF020270.
- Multinational Coordinated Arabidopsis thaliana Research Project (1997). *Progress Report: Year Six*. [Http://www.nsf.gov/pubs/1997/nsf97131/nsf97131.htm](http://www.nsf.gov/pubs/1997/nsf97131/nsf97131.htm)
- Munro, S. & Pelham, H.R. (1987): A C-terminal sequence prevents secretion of luminal ER proteins. *Cell*, **48**: 899-907.
- Murray, M.G. & Thompson, W.F. (1980): Isolation of high-molecular-weight plant DNA. *Nucl. Acids Res.*, **8**: 4321-4325.
- Nakai, K. & Kanehisa, M. (1992): A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**: 897-911.
- Nakamura, Y., Gojobori, T. & Ikemura, T. (1997): Codon usage tabulated from the international DNA sequence databases. *Nucl. Acids Res.*, **25**: 244-245.
- Needleman, S.B. & Wunsch, C.D. (1970): A general method applicable to search for similarities in the amino acid sequences of two proteins. *J.Mol.Biol.*, **48**: 443-453.
- Neupert, W. (1997): Protein Import into mitochondria. *Annu. Rev. Biochem.*, **66**: 863-917.
- Newman, T., deBruijn, F.J., Green, P., Keegstra, K., Kende, H., McIntosh, L., Ohlrogge, J., Raikhel, N., Somerville, S., Thomashow, M., Retzel, E. & Somerville, C. (1994): Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. *Plant Physiol.*, **106**: 1241-1255.

- Nicholas, K.B., Nicholas, H.B.J. & Deerfield, D.W. (1997): GeneDoc: Analysis and visualization of genetic variation. *EMBNEW.NEWS*, **4**: 14.
- O'Kane, C.J. & Gehring, W.J. (1987): Detection *in situ* of genomic regulatory elements in *Drosophila*. *Proc. Natl. Acad. Sci. USA.*, **84**: 9123-9127.
- Ohba, T., Rennert, H., Pfeifer, S.M., He, Z., Yamamoto, R., Holt, J.A., Billheimer, J.T. & Strauss, J.F.3. (1994): The structure of the human sterol carrier protein X/sterol carrier protein 2 gene (SCP2). *Genomics*, **24**: 370-374.
- Okresz, L., Mathe, C., Horvath, E., Schell, J., Koncz, C. & Szabados, L. (1998): T-DNA trapping of a cryptic promoter identifies an ortholog of a highly conserved SNZ growth arrest response genes in *Arabidopsis*. *Plant Sci.*, **138**: 213-228.
- Olesen, C., Thomsen, K.K., Svendsen, I. & Brandt, A. (1997): The glyoxysomal 3-ketoacyl-CoA thiolase precursor from *Brassica napus* has enzymatic activity when synthesized in *Escherichia coli*. *FEBS Lett.*, **412**: 138-140.
- Oppenheimer, D.G., Pollock, M.A., Vacik, J., Szymanski, D.B., Ericson, B. & Feldmann, K.A. (1997): Essential role of a kinesin-like protein in *Arabidopsis* trichome morphology. *Proc. Natl. Acad. Sci. USA*, **94**: 6261-6266.
- Osborne, B.I., Corr, C.A., Prince, J.P., Hehl, R., Tanksley, S.D., McCormick, S. & Baker, B. (1991): *Ac* transposition from a T-DNA can generate linked and unlinked clusters of insertions in the tomato genome. *Genetics*, **129**: 833-844.
- Osborne, B.I., Vysotskaia, V.S., Toriumi, M., Yu, G., Oji, O., Liu, S., Li, J., Hoang, L., Buehler, E., Conway, A.B., Conway, A.R., Dewar, K., Feng, J., Kim, C., Kurtz, D., Li, Y., Shinn, P., Sun, H., Davis, R.W., Ecker, J.R., Federspiel, N.A. & Theologis, A. (1997): The sequence of BAC T1G11 from *Arabidopsis thaliana* chromosome 1. *GenBank*. Accession AC002376 (locus T1G11).
- Palmer, M., Differding, E., Gamboni, R., Williams, S.F., Peoples, O.P., Walsh, C.T., Sinskey, A.J. & Masamune, S. (1991): Biosynthetic thiolase from *Zoogloea-ramigera* - evidence for a mechanism involving Cys-378 as the active-site base. *J. Biol. Chem.*, **266**: 8369-8375.
- Park, J.-H., Oh, S.A., Kim, Y.H., Woo, H.R. & Nam, H.G. (1998): Differential expression of senescent-associated mRNAs during leaf senescence induced by different senescence-inducing factors in *Arabidopsis*. *Plant Mol. Biol.*, **37**: 445-458.

- Passreiter, M., Anton, M., Lay, D., Frank, R., Harter, C. & Wieland F.T. (1998). Peroxisome biogenesis: involvement of ARF and coatamer. *J. Cell. Biol.* **141**: 373-383.
- Pearson, W.R. & Lipman, D.J. (1988): Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **85**: 2444-2448.
- Peitsch, M.C. (1995): Protein Modelling by E-mail. *Bio/Technology*, **13**: 658-660.
- Peitsch, M.C. (1996): ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans.*, **24**: 274-279.
- Peitsch, M.C. & Jongeneel, C.V. (1993): 3-D model for the CD40 ligand predicts that it is a compact trimer similar to the tumor necrosis factors. *Int. Immunol.*, **5**: 233-238.
- Peoples, O.P., Masamune, S., Walsh, C.T. & Sinskey, A.J. (1987): Biosynthetic thiolase from *Zoogloea ramigera*. III. Isolation and characterization of the structural gene. *J. Biol. Chem.*, **262**: 97-102.
- Peoples, O.P. & Sinskey, A.J. (1989): Poly-beta-hydroxybutyrate biosynthesis in *Alcaligenes eutrophus* H16. Characterization of the genes encoding beta-ketothiolase and acetoacetyl-CoA reductase. *J. Biol. Chem.*, **264**: 15293-15297.
- Perez-Grau, L. & Goldberg, R.B. (1989): Soybean seed protein genes are regulated spatially during embryogenesis. *Plant Cell*, **1**: 1095-1109.
- Petersen, D.J. & Bennett, G.N. (1991): Cloning of the *Clostridium acetobutylicum* ATCC 824 acetyl coenzyme A acetyltransferase (thiolase; EC 2.3.1.9) gene. *Appl. Environ. Microbiol.*, **57**: 2735-2741.
- Phillips, A.L. & Huttly, A.K. (1994): Cloning of two gibberelin-regulated cDNAs from *Arabidopsis thaliana* by subtractive hybridization: expression of the tonoplast water channel, g-TIP, is increased by GA₃. *Plant Mol. Biol.*, **24**: 603-615.
- Portugal, F.H. & Cohen, J.S. (1979). *A century of DNA*. Cambridge, Massachusetts London: MIT Press.
- Preisig-Mueller, R., Ahnert, V. & Kindl, H. (1993). Primary structure of the hsp70 of the cucumber chloroplastic stroma. GenBank. Accession CAA52149.
- Preisig-Müller, R. & Kindl, H. (1993): Thiolase mRNA translated in vitro yields a peptide with a putative N-terminal presequence. *Plant Mol. Biol.*, **22**: 59-66.

- Prestridge, D.S. (1991): A computer program that scans DNA sequences for eukaryotic transcriptional elements. *Comput. Appl. Biosci.*, **7**: 203-206.
- Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. (1995): MatInd and MatInspector-New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids Res.*, **23**: 4878-4884.
- Raghavan, V. (1986). *Embryogenesis in Angiosperms*. Cambridge: Cambridge University Press.
- Reiser, L. & Fischer, R.L. (1993): The ovule and the embryo sac. *Plant Cell*, **5**: 1291-1301.
- Richardson, J.S. & Richardson, D.C. (1989). Principles and patterns of protein conformation. In G. D. Fasman (Eds.), *Prediction of protein structure and the principles of protein conformation* (pp. 1-98). New York: Plenum Press.
- Rocha, P.S.C.F. (1991) *Analysis of the molecular organization and structure of the nuclear rDNA in Brassica rapa*. M.Sc., University of Guelph, Ontario, Canada.
- Roshe, E. & Westhoff, P. (1995). Genomic structure and expression of the pyruvate, orthophosphate dikinase gene of the dicotyledonous C-4 plant *Flaveria trinervia* (Asteraceae). *Plant Mol. Biol.* **29**: 663-678.
- Rost, B. & Sander, C. (1993): Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**: 584-599.
- Rost, B. & Sander, C. (1994): Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**: 55-77.
- Rost, B., Sander, C. & Schneider, R. (1994): PHD-an automatic mail server for protein secondary structure prediction. *Comput. Appl. Biosci.*, **10**: 53-60.
- Rounsley, S.D., Ketchum, K.A., Lin, X., Crosby, M.L., Brandon, R.C., Sykes, S.M., Kaul, S., Mason, T.M., Kerlavage, A.R., Adams, M.D., Sommerville, C.R. & Venter, J.C. (1997): *Arabidopsis thaliana* chromosome II BAC F25I18 genomic sequence. GenBank. Accession AC002334 (locus ATAC002334).
- Rushton, P., Torres, J., Parniske, m., Wernert, P., Hahlbrock, K. & Somssich, I. (1996): Interaction of elicitor-induced DNA-binding proteins with elicitor response elements in the promoters of parsley PR1 genes. *EMBO J.*, **15**: 5690-5700.
- Salisbury, F.B. & Ross, C.W. (1993). *Plant Physiology*. Belmont, California: Wadsworth Publishing Company.

- Sambrook, J., Fritsch, E.F. & Maniatis, T. (1989). *Molecular cloning. A laboratory manual*. Cold Spring Harbor: Cold Spring Laboratory Press.
- Sanger, F., Nicklen, S. & Coulson, A.R. (1977): DNA sequencing with chain-terminating inhibitors.
Proc. Natl. Acad. Sci. USA, **74**: 5463-5467.
- Sato, S., Hayashi, M., Imamura, S., Ozeki, Y. & Kawaguchi, A. (1992): Primary structures of the genes, *faoA* and *faoB*, from *Pseudomonas fragi* B-0771 which encode the two subunits of the HDT multienzyme complex involved in fatty acid beta-oxidation.
J. Biochem., **111**: 8-15.
- Sawadogo, M. & van Dike, M.W. (1991): A rapid method for the purification of deprotected oligodeoxynucleotides.
Nucl. Acids Res., **19**: 674.
- Scheres, B., Wolkenfelt, H., Willemsen, V., Terlouw, M., Lawson, E., Dean, C. & Weisbeek, P. (1994): Embryonic origin of the *Arabidopsis* primary root and root meristem initials.
Development, **120**: 2475-2487.
- Schiavone, F.M. & Racusen, R.H. (1991): Regeneration of the root pole in surgically transected carrot embryos occurs by position-dependent, proximal-distal replacement of missing tissues.
Development, **113**: 1305-1313.
- Schneider, T., Dinkins, R., Robinson, K., Shellhammer, J. & Meinke, D.W. (1989): An embryo-lethal mutant of *Arabidopsis thaliana* is a biotin auxotroph.
Dev. Biol., **131**: 161-167.
- Schuler, G.D., Altschul, S.F. & Lipman, D.J. (1991): A workbench for multiple alignment construction and analysis.
Proteins: Structure, Function and Genetics, **9**: 180-190.
- Schultz, C.J. & Coruzzi, G.M. (1995). The aspartate aminotransferase gene family of *Arabidopsis* encodes isoenzymes localized to three distinct subcellular compartments.
Plant J. **7**: 61-75.
- Scott, M.R.D., Westphal, K.-H. & Rigby, P.W.J. (1983): Activation of mouse genes in transformed cells.
Cell, **34**: 557-567.
- Sedgwick, S.G. & Morgan, B.A. (1994): Locating, DNA sequencing, and disrupting yeast genes using *Tn1000*.
Meth. Mol. Genet., **3**: 131-140.
- Seedorf, U., Raabe, M. & Assmann, G. (1993): Cloning, expression and sequences of mouse sterol-carrier protein-x-encoding cDNAs and a related pseudogene.
Gene, **123**: 165-172.

- Shcheban, T.Y., Shi, J., Durachko, D.M. & Guiltinan, M.J. (1995): Molecular cloning and sequence analysis of expansins-a highly conserved, multigene family of proteins that mediate cell walls extension in plants.
Proc. Natl. Acad. Sci. USA, **92**: 9245-9249.
- Sheen, J. (1991). Molecular mechanisms underlying the differential expression of maize pyruvate, orthophosphate dikinase genes.
Plant Cell **3**: 225-245.
- Sheridan, W.F. & Clark, J.K. (1993): Mutational analysis of morphogens of the maize embryo.
Plant J., **3**: 347-358.
- Shevell, D.E., Leu, W.-M., Gillmour, C.S., Xia, G., Feldmann, K.A. & Chua, N.-H. (1994): *EMB30* is essential for normal cell division, cell expansion, and cell adhesion in Arabidopsis and encodes a protein that has similarity to Sec7.
Cell, **77**: 1051-1062.
- Shirsat, A., Wilford, N., Croy, R. & Boulter, D. (1989): Sequences responsible for the tissue specific promoter activity of a pea legumin gene in tobacco.
Mol. Gen. Genet., **215**: 326-331.
- Short, J.M., Fernandez, J.M., Sorge, J.A. & Huse, W.D. (1988): lZAP: a bacteriophage λ expression vector with *in vivo* expression properties.
Nucl. Acids Res., **16**: 7583-7600.
- Sieburth, L.E. & Meyerowitz, E.M. (1997): Molecular dissection of the *AGAMOUS* control region shows that *cis* elements for spatial regulation are located intragenically.
Plant Cell, **9**: 355-365.
- Simpson, C.G., Leader, D.J., Brown, J.W.S. & Franklin, T. (1993). Characteristics of plant pre-mRNA introns and transposable elements. In R. R. D. Croy (Eds.), *Plant Molecular Biology LABFAX* (pp. 183-251). Bios Scientific Publishers Limited.
- Simpson, C.G., Simpson, G.G., Clark, G., Leader, D.J., Vaux, P., Guerineau, F., Waugh, R. & Brown, J.W.S. (1992): Splicing of plant pre-mRNAs.
Proc. Roy. Soc. Edinburgh, **99B**: 31-50.
- Skene, M. (1939). *The biology of flowering plants*. London, U.K.: Sidgwick & Jackson, Ltd.
- Slack, J.M.W. (1983). *From egg to embryo*. Cambridge: Cambridge University Press.
- Smith, G.P. (1985): Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface.
Science, **228**: 1315-1317.
- Song, X.Q., Fukao, T., Yamaguchi, S., Miyazawa, S., Hashimoto, T. & Orii, T. (1994): Molecular cloning and nucleotide sequence of complementary DNA for human hepatic cytosolic acetoacetyl-coenzyme A thiolase.
Biochem. Biophys. Res. Commun., **201**: 478-485.

- Springer, P.S., McCombie, W.R., Sundaresan, V. & Martienssen, R.A. (1995): Gene trap tagging of *PROLIFERA*, an essential *MCM2-3-5*-like gene in *Arabidopsis*. *Science*, **268**: 877-880.
- Stacey, N.J., Roberts, K. & Knox, J.P. (1990): Patterns of expression of the JIM4 arabinogalactan-protein epitope in cell culture and during somatic embryogenesis in *Daucus carota*. *Planta*, **180**: 285-292.
- Stalberg, K., Ellerstrom, M., Ezcurra, I., Ablov, S. & Rask, L. (1996): Disruption of an overlapping E-box/ABRE motif abolished high transcription of the *napA* storage-protein promoter in transgenic *Brassica napus* seeds. *Planta*, **199**: 515-519.
- Steeves, T.A. & Sussex, I.M. (1989). *Patterns in Plant Development* (2nd ed.). Cambridge: Cambridge University Press.
- Straus, D. & Ausubel, F.M. (1990): Genomic subtraction to clone DNA corresponding to deletion mutants. *Proc. Natl. Acad. Sci. USA*, **87**: 1889-1893.
- Stubbe, H. (1965). *History of Genetics*. Cambridge, Massachusetts: MIT Press.
- Sturtevant, A.H. (1965). *A history of genetics*. New York: Harper & Row, Publishers.
- Subramani, S. (1993): Protein import into peroxisomes and biogenesis of the organelle. *Ann. Rev. Cell Biol.*, **9**: 445-478.
- Subramani, S. (1996): Protein translocation into peroxisomes. *J. Biol. Chem.*, **271**: 32483-32486.
- Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R.K., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R. & Waterston, R. (1992): The *C. elegans* genome sequencing project: a beginning. *Nature*, **356**: 37-41.
- Swinkels, B.W., Gould, S.J., Bodnar, A.J., Rachubinski, R.A. & Subramani, S. (1991): A novel cleavable peroxisomal targeting signal at the amino-terminus of the rat 3-ketoacyl-CoA thiolase. *EMBO J.*, **10**: 3255-3262.
- Szekeres, M., Nemeth, K., Koncz-Kalman, Z., Mathur, J., Kauschmann, A., Altmann, T., Redei, G.P., Nagy, F., Schell, J. & Koncz, C. (1996): Brassinosteroids rescue the deficiency of CYP90, a cytochrome P450, controlling cell elongation and de-etiolation in *Arabidopsis*. *Cell*, **85**: 171-182.
- Takahashi, T., Gasch, A., Nishizawa, N. & Chua, N.-H. (1995): The *DIMINUTO* gene of *Arabidopsis* is involved in regulating cell elongation. *Genes Devel.*, **9**: 97-107.

- Tanksley, S.D., Ganai, M.W. & Martin, G.B. (1995): Chromosome landing: a paradigm for map-based cloning in plants with large genomes. *Trends in Genetics*, **11**: 63-68.
- Tatusov, R.L., Mushegian, A.R., Bork, P., Brown, N.P., Hayes, W.S., Borodovsky, M., Rudd, K.E. & Koonin, E.V. (1996): Metabolism and evolution of *Haemophilus influenzae* deduced from a whole genome comparison with *Escherichia coli*. *Curr. Biol.*, **291**:
- Taylor, C.B. (1997): Promoter fusion analysis: an insufficient measure of gene expression. *Plant Cell*, **9**: 273-275.
- Taylor, W.R. (1986): The classification of amino-acid conservation. *J. Theor. Biol.*, **119**: 205-205.
- Teeri, T.H., Herrera-Estrella, L., Depicker, A., van Montagu, M. & Palva, E.T. (1986): Identification of plant promoters *in situ* by T-DNA-mediated transcriptional fusions to the *npt II* gene. *EMBO J.*, **5**: 1755-1760.
- Terzaghi, W.B. & Cashmore, A.R. (1995): Light-regulated transcription. *Annual Rev. Plant Physiol. Mol. Biol.*, **46**: 445-474.
- Thomas, T.L. (1993): Gene expression during plant embryogenesis and germination: an overview. *Plant Cell*, **5**: 1401-1410.
- Thompson, S., Mayer, F., Peoples, O.P., Masamune, S., Synskey, A.J. & Walsh, C.T. (1989): Mechanistic studies on b-ketoacyl thiolase of *Zoogloea ramigera*: Identification of the active site nucleophile as Cys89, its mutation to Ser89, and kinetic and thermodynamic characterization of wildtype and mutant enzymes. *Biochemistry*, **28**: 5735-5742.
- Titorenko, V.I., Smith, J.J., Szilard, R.K. & Rachubinski, R.A. (1998): Pex20p of yeast *Yarrowia lipolytica* is required for the oligomerization of thiolase in the cytosol and for its targeting to the peroxisome. *J. Cell Biol.*, **142**: 403-420.
- Tolbert, N.E. (1981): Metabolic pathways in peroxisomes and glyoxysomes. *Ann. Rev. Biochem.*, **50**: 133-157.
- Tombolini, R., Povolo, S., Buson, A., Squartini, A. & Nuti, M.P. (1995): Poly-beta-hydroxybutyrate (PHB) biosynthetic genes in *Rhizobium meliloti*. *Microbiology*, **141**: 2553-2559.
- Topping, J.F., Agyeman, F., Henricot, B. & Lindsey, K. (1994): Identification of molecular markers of embryogenesis in *Arabidopsis thaliana* by promoter trapping. *Plant J.*, **5**: 895-903.

- Topping, J.F. & Lindsey, K. (1997): Promoter trap markers differentiate structural and positional components of polar development in *Arabidopsis*.
Plant Cell, **9**: 1713-1725.
- Topping, J.F., May, V.J., Muskett, P.R. & Lindsey, K. (1997): Mutations in the *HYDRA1* gene of *Arabidopsis thaliana* perturb cell shape and disrupt embryonic and seedling morphogenesis. Development,
- Topping, J.F., Wei, W. & Lindsey, K. (1991): Functional tagging of regulatory elements in the plant genome.
Development, **112**: 1009-1019.
- Triglia, T., Peterson, M.G. & Kempt, D.J. (1988): A procedure for *in vitro* amplification of DNA segments that lie outside known sequences.
Nucl. Acids Res., **16**: 8186-8186.
- Truelsen, E., Gausing, K., Joachimsen, B., Jorgensen, P. & Marcker, K.A. (1979): Cloning of soybean leghaemoglobin structural gene sequences synthesized *in vitro*.
Nucl. Acids Res., **6**: 3061-3072.
- Tsugeki, R., Kochieva, E.Z. & Fedoroff, N.V. (1996): A transposon insertion in the *Arabidopsis* *SSR15* gene causes an embryo-defective lethal mutation.
Plant J., **10**: 479-489.
- Tsui, L.-C. & Estivill, X. (1991). Identification of disease genes on the basis of chromosomal localization. In K. E. Davies & S. M. Tilghman (Eds.), *Genes and phenotypes* (pp. 1-36). Cold Spring Harbor: Cold Spring Harbor Laboratory Press.
- Uberbacher, E.C. & Mural, R.J. (1991): Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach.
Proc. Natl. Acad. Sci. USA, **88**: 11261-11265.
- Ulmasov, T., Hagen, G. & Guilfoyle, T.J. (1997): ARF1, a transcription factor that binds to auxin response elements.
Science, **276**: 1865-1868.
- Umeda, M., Bhalerao, R.P., Schell, J., Uchimiya, H. & Koncz, C. (1998): A distinct cyclin-dependant kinase-activating kinase of *Arabidopsis thaliana*.
Proc. Natl. Acad. Sci. USA, **95**: 5021-5026.
- Uwer, U., Willmitzer, L. & Altmann, T. (1998): Inactivation of a glycyl-tRNA synthetase leads to an arrest in plant embryo development.
Plant Cell, **10**: 1277-1299.
- van Camp, W., Bowler, C., Villaroel, R., Tsang, E.W.T., Van Montagu, M. & Inze, D. (1990): Characterization of iron superoxide dismutase cDNAs from plants obtained by genetic complementation of *Escherichia coli*.
Proc. Natl. Acad. Sci. USA, **87**: 9903-9907.

- van den Berg, C., Willemsen, V., Hage, W., Weisbeek, P. & Scheres, B. (1995): Cell fate in the *Arabidopsis* root meristem determined by directional signalling. *Nature*, **378**: 62-65.
- van Nocker, S., Deveraux, Q., Rechsteiner, M. & Vierstra, R.D. (1996): *Arabidopsis MBP1* gene encodes a conserved ubiquitin recognition component of the 26S proteasome. *Proc. Natl. Acad. Sci. USA*, **93**: 856-860.
- van Sluys, M.A., Tempe, J. & Fedoroff, N. (1987): Studies on the introduction and mobility of the maize *Activator* element in *Arabidopsis thaliana* and *Daucus carota*. *EMBO J.*, **6**: 3881-3889.
- van Straeten, D., van Wiemeersch, L., Goodman, H.M. & van Montagu, M. (1992): Cloning and sequencing of two different cDNAs encoding 1-aminocyclopropane-1-carboxylate synthase in tomato. *Proc. Natl. Acad. Sci. USA*, **87**: 4854-4863.
- Vaucheret, H., Beclin, C., Elmayan, T., Feuerbach, F., Godon, C., Morel, J.-B., Mourrain, P., Palauqui, J.-C. & Vernhettes, S. (1998): Transgene-induced gene silencing in plants. *Plant J.*, **16**: 651-659.
- Vernon, D.M. & Meinke, D.W. (1994): Embryogenic transformation of the suspensor in *twin*, a polyembryonic mutant of *Arabidopsis*. *Dev. Biol.*, **165**: 566-573.
- Vollack, K.U. & Bach, T.J. (1996): Cloning of a cDNA-encoding cytosolic acetoacetyl-coenzyme-A thiolase from radish by functional expression in *Saccharomyces cerevisiae*. *Plant Physiol.*, **111**: 1097-1107.
- Vos, P., Hogers, R., M., B., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kulper, M. & Zabeau, M. (1995): AFLP: A new technique for DNA fingerprinting. *Nucl. Acids Res.*, **23**: 4407-4414.
- Walbot, V. (1992): Strategies for mutagenesis and gene cloning using transposon tagging and T-DNA insertional mutagenesis. *Ann. Rev. Plant Physiol. Plant Mol. Biol.*, **47**: 49-82.
- Walden, R., Fritze, K., Hayashi, H., Miklashevichs, E., Harling, H. & Schell, J. (1994): Activation tagging: a means of isolating genes implicated as playing a role in plant growth and development. *Plant Mol. Biol.*, **26**: 1521-1528.
- Wang, M.M. & Reed, R.R. (1993): Molecular cloning of the olfactory neuronal transcriptinal factor Olf-1 by genetic selection in yeast. *Nature*, **364**: 121-126.
- Waterham, H.R., Titorenko, V.I., Haima, P. & Cregg, J.M. (1994): The *Hansenula polymorpha* PER1 gene is essential for peroxisome biogenesis and encodes a peroxisomal matrix protein with both carboxy- and amino-terminal targeting signals. *J. Cell Biol.*, **127**: 737-749.

- Wei, W. (1994): Promoter-trapping vector pDeltagusBin19. GenBank. Accession U12638.
- Wei, W., Twell, D. & Lindsey, K. (1997): A novel nucleic acid helicase gene identified by promoter-trapping in *Arabidopsis*. *Plant J.*, **11**: 1307-1314.
- West, M.A.L. & Harada, J.J. (1993): Embryogenesis in higher plants: an overview. *Plant Cell*, **5**: 1361-1369.
- Whitehouse, H.L.K. (1973). *Towards an understanding of the mechanism of heredity* (3rd Edition ed.). London: Edward Arnold Publishers, Ltd.
- Wilkins, A.S. (1986). *Genetic analysis of animal development*. New York, Toronto, Singapore: John Wiley & Sons.
- Wilkinson, J.Q., Lanahan, M.B., Conner, T.W. & Klee, H.J. (1995): Identification of mRNAs with enhanced expression in ripening strawberry fruit using polymerase chain reaction differential display. *Plant Mol. Biol.*, **27**: 1097-1108.
- Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A. & Tingey, S.V. (1990): DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucl. Acids Res.*, **18**: 6531-6535.
- Williams, S.F., Palmer, M.A.J., Peoples, O.P., Walsh, C.T., Sinskey, A.J. & Masamune, S. (1992): Biosynthetic thiolase from *Zoogloea ramigera*. Mutagenesis of putative active-site base Cys378 changes the partitioning of the acetyl S-enzyme intermediate. *J. Biol. Chem.*, **267**: 16041-16043.
- Wimmer, B., Lottspeich, F., van der Klei, Veenhuis, M. & Gietl, C. (1997). The glyoxysomal and plastid molecular chaperones (70-kDa heat-shock protein) of watermelon cotyledons are encoded by a single gene. *Proc. Natl. Acad. Sci. USA*, **94**: 13624-13629.
- Witt, U., Rehberg, S. & Abel, W.O. (1995): A Full-Length cDNA (Accession No. X92512) coding for glyoxysomal malate dehydrogenase from *Brassica napus* L. *Plant Physiol.*, **110**: 336-336.
- Wu, Y., Haberland, G., Zhou, C. & Koop, H.-U. (1992): Somatic embryogenesis, formation of morphogenetic callus and normal development in zygotic embryos of *Arabidopsis thaliana*. *Protoplasma*, **169**: 89-96.
- Xue, Y., Mural, R.J., Shah, M.B. & Uberbacher, E.C. (1994): Recognizing exons in genomic sequence using GRAIL II. *Genetic Engineering: principles and methods.*, **16**: 241-253.

- Yadegari, R., de Paiva, G.R., Laux, T., Koltunow, A.M., Apuya, N., Zimmerman, J.L., Fischer, R.L., Harada, J.J. & Goldberg, R.B. (1994): Cell differentiation and morphogenesis are uncoupled in *Arabidopsis raspberry* embryos.
Plant Cell, **6**: 1713-1729.
- Yanisch-Perron, C.J., Vieira, J. & Messing, J. (1985): Improved M13 phage cloning vectors and host strains: Nucleotide sequences of the M13mp18 and pUC19 vectors.
Gene, **33**: 103-119.
- Yoshioka, S., Kato, K., Nakai, K., Okayama, H. & Nojima, H. (1997): Identification of open reading frames in *Schizosaccharomyces pombe* cDNAs.
DNA Res., **4**: 363-369.
- Zambryski, P., Tempe, J. & Schell, J. (1989): Transfer and function of T-DNA genes from *Agrobacterium* Ti and Ri plasmids in plants.
Cell, **56**: 193-201.
- Zhang, J.F. & Somerville, C.R. (1997): Suspensor derived polyembryony caused by altered expression of valyl-tRNA synthetase in the *town2* mutant of *Arabidopsis*.
Proc. Natl. Acad. Sci. USA, **94**: 7349-7355.
- Zhou, D., Li, Y., Rocipon, M. & Mache, R. (1992): Sequence specific interaction between S1F, a spinach nuclear factor, and a negative cis-element conserved in plastid-related genes.
J. Biol. Chem., **267**: 23515-23519.
- Zimmerman, J.L. (1993): Somatic embryogenesis: a model for early development of higher plants.
Plant Cell, **5**: 1411-1423.