PHD

Isolation of novel expressed DNA sequences from cassava (Manihot esculenta, Crantz)

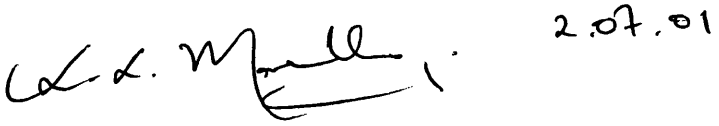
Marello, Karen Louise

Award date:
2000

Awarding institution:
University of Bath

ISOLATION OF NOVEL EXPRESSED DNA SEQUENCES FROM CASSAVA (MANIHOT ESCULENTA, CRANTZ)

submitted by Karen Louise Marello
for the degree of PhD
of the University of Bath
2000



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

CONTENTS **Page**

ABSTRACT

Nutrient studies using differing amounts of nitrogen and sulphur were carried out on cassava plants to investigate the effect on storage root production. Unexpectedly, the combination of high levels of both additional nutrients dramatically reduced root yield, with one plant having no storage roots.

A cDNA library was constructed from cassava storage root parenchyma tissue, and screened by a number of methods including the polymerase chain reaction, which yielded several clones. These were sequenced, identified and four selected for further analysis.

A cassava homologue for the nascent polypeptide associated complex and activator α chain (α-NAC) sequence was obtained. This had a potential open reading frame of 208 amino acids, was found to be 73% identical to a putative α-NAC from *Arabidopsis thaliana* and is apparently part of a multi-gene family.

Two cassava partial ribosomal DNA clones were isolated and used for phylogenetic studies. One represented 1169bp of the 26S subunit and was compared with the same region from eight other plant 26S DNA sequences and one algal sequence. The other represented 686bp of the 18S subunit DNA sequence and was compared with the same region from 40 plant and one algal 18S DNA sequences.

The final cDNA to be studied was an apparently novel cassava sequence, which appeared from Southern analysis to be present as a single-copy gene. cDNA sequence corresponding to an open reading frame of 166 amino acid residues was obtained, with a helix-loop-helix motif and a potential coil region. The corresponding genomic clone was obtained and sequenced for 3670bp. No identity could be assigned to this cassava sequence based on comparison with other DNA sequences, although it may represent a DNA-binding protein. Various regulatory elements have been suggested, based on comparison with genes from cassava and other plants.

ACKNOWLEDGEMENTS

My thanks to my supervisors, Dr John Beeching and Professor Peter Shewry, for their confidence in me, their enthusiasm and their constructive comments on this thesis.

To Mike, who looked after the cassava plants in the greenhouse and reminded me when things needed tending, and Rob, who usually had the job of spraying the pesticides – thanks! Thanks also to Madan, who encouraged me to think positively about my results and to Tim who did all the automated sequencing. I am also very grateful to Roger Guevara, Dr Roger Mepsted and Dr Mike Mogie who all gave me advice on statistics at one time or another.

A big thank you to all my friends and colleagues at the university over the years I spent at Bath, especially Lab 1.52. In particular: to Janet, for many a shopping spree when it all got too much; to Paul, whose wry, dry comments were so frequently in tune with my own; to compañera Virginia, who introduced us to caipirinha, and shared many a fiesta; and Tasha, otra salsera, without whose company for dancing, partying and late-night-lab-sessioning I'd have run away screaming!!!

Finally, my thanks, my appreciation and my love to three people: Sylvia and Bob, my parents, who always encouraged me, supported me and respected my decision to defer gainful employment and storage space for my possessions for yet another few years; and Jeremy, without whom, quite simply, this thesis would not be.

Thank you all.

List of abbreviations

Adenosine diphosphate glucose pyrophosphorylase	AGPase
African cassava mosaic virus	ACMV
Bacteriophage lambda	λ
Base pair	bp
Bovine serum albumin	BSA
Branching enzyme	BE
5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside	X-gal
Cassava Biotechnology Network	CBN
Centro Internacional de Agricultura Tropical	CIAT
Complementary DNA	cDNA
Day(s)	d
Deoxyadenosine-; deoxycytosine-; deoxyguanosine-;	dATP; dCTP; dGTP;
deoxynucleotide-; and deoxythymidine-triphosphate	dNTP; dTTP
Deoxynucleotide monophosphate	dNMP
Deoxyribonuclease	DNase
Deoxyribonucleic acid	DNA
Deoxyribonucleotide triphosphate	dNTP
Diaminobenzidine	DAB
Di-deoxy-adenosine; -cytosine; -guanosine; -thymidine	ddA; ddC; ddG; ddT
Diethylpyrocarbonate	DEPC
Dimethylsulphoxide	DMSO
Dithiothreitol	DTT
East African cassava mosaic virus	EACMV
Electrochemiluminescence	ECL
Ethylenediamine tetra-acetic acid – disodium salt	EDTA
European Molecular Biology Laboratory	EMBL
Expressed sequence tag	EST
External transcribed spacer	ETS
Far upstream element	FUE
Food and Agricultural Organisation	FAO
Friable embryogenic callus	FEC
Genomic DNA	gDNA
Granule bound starch synthase	GBSS
Human Genome Mapping Project	HGMP
Hydroxyproline-rich glycoproteins	HPRGs
Hour	h
Immunoglobulin	Ig
Indian cassava mosaic virus	ICMV
Intergenic spacer	IGS
Internal transcribed spacer	ITS
International Institute of Tropical Agriculture	IITA
Isoamylalcohol	IAA
Isopropyl β-D-thiogalactopyranoside	IPTG
β-glucoronidase	GUS
Kilo base	kb
Kilo Dalton	kDa

List of abbreviations – continued

Luria-Bertani medium	LB
Luria-Bertani medium plus maltose	LBM
Messenger RNA	mRNA
Minute	min
Namulonge Agricultural & Animal Research Institute	NAARI
Nascent polypeptide associated complex and activator	NAC
Near upstream element	NUE
3-(N-morpholino) propanesulphonic acid	MOPS
N, N, N', N'-tetramethylethylenediamine	TEMED
Open reading frame	ORF
Phenylalanine ammonia lyase	PAL
Polyacrylamide gel electrophoresis	PAGE
Polyethylene glycol	PEG
Polymerase chain reaction	PCR
Post-harvest physiological deterioration	PPD
Pounds per square inch	psi
Random amplified polymorphic DNA	RAPD
Rapid amplification of cDNA ends	RACE
Restriction fragment length polymorphism	RFLP
Reverse transcriptase PCR	RT-PCR
Revolutions per minute	rpm
Ribonuclease	RNase
Ribonucleic acid	RNA
Ribosomal DNA	rDNA
Ribosomal RNA	rRNA
Salt-magnesium buffer	SM
Salt-Tris-EDTA buffer	STE
Sodium dodecyl sulphate	SDS
Soluble starch synthase	SSS
Species	*spp.*
Standard sodium citrate	SSC
Sterile MilliQ water	SMQ
Thermophilus aquaticus	*Taq*
Transcription start site	TCS
Transformation and storage solution	TSS
Tris-Acetate-EDTA	TAE
Tris-Borate-EDTA	TBE
Tris-buffered saline	TBS
Tris-buffered saline plus Tween 20	TBST
Tris-EDTA	TE
Tris(hydroxymethyl)aminomethane hydrochloride	Tris-HCl
Ultraviolet	UV
Untranslated region	UTR
Upstream binding factor	UBF
Uridine diphosphate	UDP
Volume per volume	v/v
Weight per volume	w/v

CHAPTER 1: GENERAL INTRODUCTION

1.1 Food security and cassava

Food is fundamental to life, yet at the end of the twentieth century the guarantee of an adequate daily supply of nutrients is non-existent for millions of people worldwide. The issue of food security is important and complex. In 1999, the world population reached 6 billion, with projections of an annual increase of 73 million people until 2020 (Pinstrup-Anderson *et al.*, 1999). Since the so-called 'Green Revolution' of the 1960s, grain production has doubled and livestock production trebled. These increases have resulted in the theoretical availability of 2700 calories per person per day, throughout the world. This is slightly above the recommended daily intake in Western Europe of 2000 and 2500 calories for a healthy adult woman and man, respectively. However, despite this achievement, around 820 million people are estimated to lack access to sufficient food, most of whom farm small plots of land in the tropics (Persley and Doyle, 1999; Pinstrup-Anderson *et al.*, 1999). Many factors contribute to this unequal distribution of food. Obvious factors such as poverty, war, drought and inappropriate land use are some of the direct causes, but political ambitions, corruption and control of global economies lie behind many of these. Feeding the world is stated by many agricultural researchers as their ultimate goal, but this will only be achieved by concerted regional, national and international efforts identifying the most appropriate areas of investment for the world's poor.

Cassava (*Manihot esculenta*, Crantz) is viewed as a crop with the potential to provide food security in many areas of the tropical world, particularly in Sub-Saharan Africa which has regions of extreme poverty. A root crop which produces secondarily thickened roots packed full of starch, cassava feeds more than 500 million people, with

almost 166 million metric tonnes produced globally in 1999 (FAO, 2000). Reliance on cassava for food energy is heaviest in Africa, where the ten highest cassava-consuming nations are found, but it is also grown in Latin America and Asia. Today, the major producers of cassava in these three continents are Nigeria and the Democratic Republic of Congo (formerly Zaire) in Africa, Brazil and Colombia in the Americas, and Indonesia and Thailand in Asia. In Africa, production and consumption of the crop are tightly linked, while in Asia much of that grown is for export. In 1996, world production of cassava exceeded 162 million tonnes of fresh roots, around 83 million tonnes of which were grown in Africa (FAO, 2000).

In terms of calories, cassava is the fourth most important crop in the world's tropical regions (after rice, maize and sugarcane) and has the best yield per hectare of all staple crops. Approximately 35% of the root is dry matter, with up to 85% of this being starch. It is cheap, vegetatively propagated by stem cuttings and requires low labour input. Furthermore, cassava will tolerate poor soils and is resistant to drought, all of which make it ideal to be grown by small-scale farmers in adverse conditions. The harvest of many crops is restricted to a particular time of year, but this is not true of cassava and it has the advantage that the roots, which are not genuine tubers or propagules, can be partially harvested while the remainder is left in the ground until required. Another advantage in areas prone to locust swarms is that cassava can withstand defoliation, with around 80% yield being maintained. This ability also allows for deliberate removal of leaves by farmers if plants are infected, without loss of the roots. Thus cassava is potentially one of the most efficient carbohydrate sources and an excellent famine reserve (reviews, Cock, 1985; Bertram, 1991; Puonti-Kaerlas, 1998).

Figure 1.1 shows a field of cassava growing in Uganda, while Figure 1.2 shows the harvested roots from a single plant.

Figure 1.1 A field in Uganda grown with the tropical shrub cassava.

Figure 1.2 The starchy 'tuberous' roots for which cassava is primarily cultivated. These have been harvested from a single Ugandan plant.

1.2 Biotechnological applications and the launch of the CBN

Large sums of money are invested in food research each year, but these are mainly from areas such as Western Europe and North America, where food is abundant. The funding for projects aimed at improving food security elsewhere is limited and needs to be appropriate.

Funding for research on cassava has been minimal, with the result that the studies carried out have been, until recently, mainly observational and field-based. Compared with other crops of major world importance, it is still massively under-funded. However, in 1988, the Cassava Biotechnology Network (CBN) was set up, to identify cassava research needs and apply biotechnology to them. Fundamental to this procedure has been the creation of links between researchers and their institutions, and the cassava farmers, processors and end-users. For example, workshops were held in Kenya with a representative group of resource-poor farmers to assess research proposals, while in Colombia the Comite Campesino Regional, consisting of farmers, meets frequently and has two representatives on the CBN Steering Committee (as cited in van den Bos, 1997).

Biology has traditionally been one of the easiest disciplines to justify, with its direct and obvious applications to human health, welfare and lifestyle, but currently there are both national and international objections to almost all of the fastest moving areas of research. Biotechnology has become a term deliberately used by the media to create panic and horror, rather than being portrayed as a means to solve problems. Ethical issues such as gene-transfer, cloning and possible effects on the environment are extremely controversial, and central to much of the current media and public interest are the cultivation, and particularly the modification, of food, for both humans and the animals they eat.

This debate over the benefits of biotechnology, rather than accepting them automatically, is particularly relevant to a crop such as cassava. Cultivated only in

tropical areas, where people are often poor and inexperienced in biotechnological issues, the question of whether it is worth applying the biotechnological possibilities developed mainly in Western European and North American labs must be answered honestly, to avoid wasting precious resources. Simply because the technology is possible does not make it desirable.

Several areas were identified by the CBN within cassava research where it was considered that biotechnology could make a significant contribution. Despite its many advantages, cassava has several limitations and is beset by a number of serious pests and diseases. Additionally, there are areas where the economic value of cassava could be enhanced, e.g. the composition of the root, particularly in terms of starch and protein. More general applications of biotechnology include genome studies, and the development of methods to improve the production and preservation of tissue. The main areas currently under investigation are detailed below.

1.2.1 The storage root – a source of starch

The main reason for the cultivation of cassava is the large 'tuberous' storage roots (Figure 1.2) which, as mentioned above, are predominantly filled with starch. Starch is the form in which carbohydrate is stored in the plant, in organelles called amyloplasts, located within the parenchyma tissue. Both chloroplasts and amyloplasts are sites of starch biosynthesis, the former having a high turnover of starch, while the latter store the plant's 'reserve' starch. Figure 1.3 shows a transverse section through a young cassava root, with enlargements showing starch granules within the parenchyma. Figure 1.4 details the enzymatic pathway from the hexose substrate to amylopectin and amylose, the two components of starch.

Figure 1.3 Transverse section through a young cassava root, highlighting the storage of starch granules within the parenchyma tissue (taken from Hunt *et al.*, 1977).

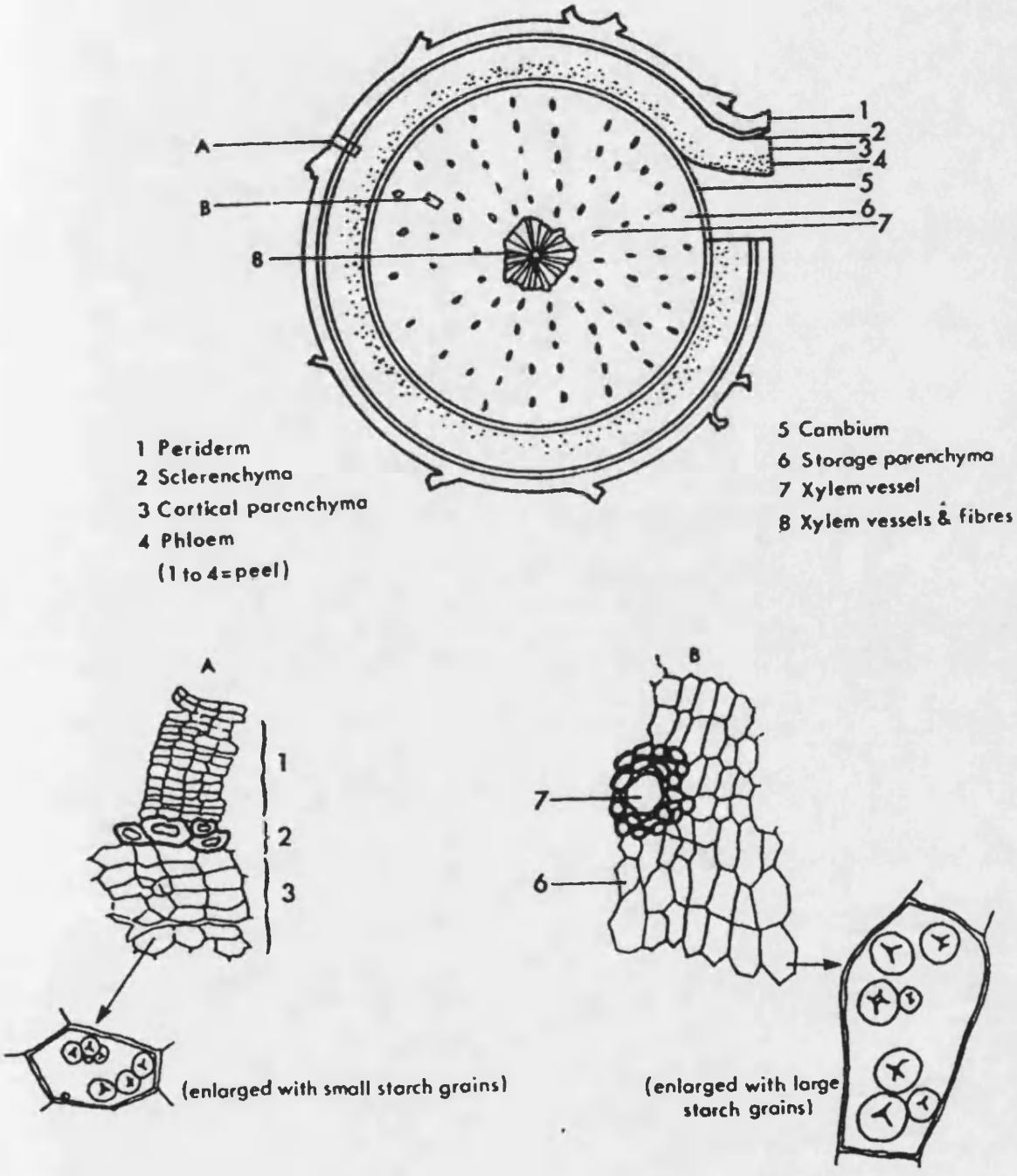
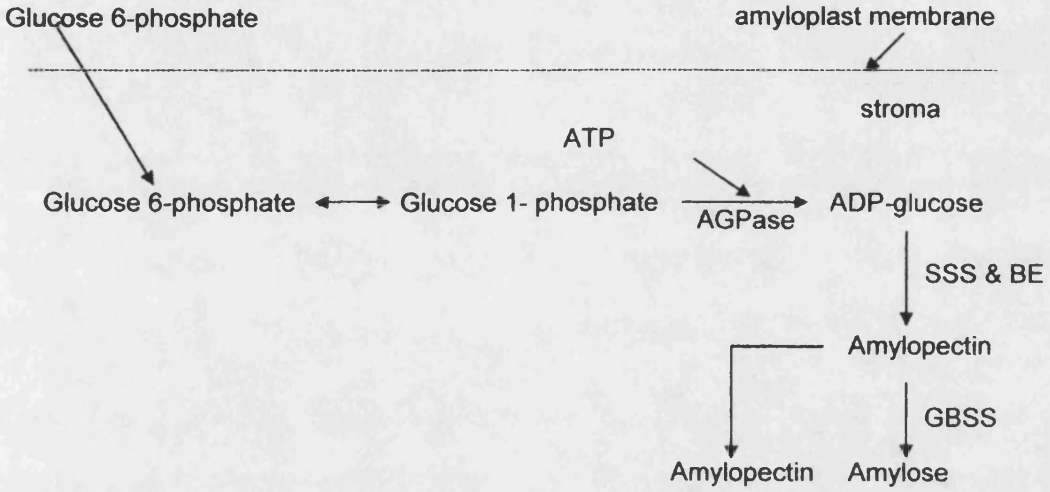


Figure 1.4 Biosynthetic pathway showing the production of the two polymers of which starch is composed, amylopectin and amylose, in the amyloplast. AGPase: adenosine diphosphate glucose pyrophosphorylase; SSS: soluble starch synthase; BE: branching enzyme; GBSS: granule-bound starch synthase (after Salehuzzamen *et al.*, 1992; Emes and Tobin, 1993; van de Wal *et al.*, 1998).



Both amylopectin and amylose are polymers, made of units of D-glucose linked by α-1,4 glycosidic bonds. Amylopectin is crosslinked with α-1,6 glycosidic bonds also, while amylose is linear and longer than amylopectin, consisting of between 100 and 1000 units. Starch generally comprises 75–80% amylopectin and 20–25% amylose. As shown in Figure 1.4, amylopectin production occurs first, catalysed by the soluble form of starch synthase together with branching enzyme which synthesises the α-1,6 crosslinks. Production of amylose requires the extension of amylopectin, followed by cleavage to produce the linear amylose molecules (van de Wal *et al.*, 1998). This reaction requires the granule-bound starch synthase (GBSS) enzyme, located bound to the starch granule. These three enzymes exist as multiple isoforms in most plants and so far a single branching enzyme isoform, two granule-bound starch synthases and three AGPases have been cloned from cassava (Salehuzzaman *et al.*, 1992, 1993; Munyikwa *et al.*, 1994, 1997). The composition of starch, particularly this ratio of

amylopectin: amylose as described above, is responsible for the properties of starch and hence its uses.

Cassava starch is produced in many areas of South East Asia and Latin America, for use in both the food and the non-food industries. While many of the producers are small-scale, this is not exclusively so, and cassava represents significant cash income for many rural and indeed national economies. For example, the Union of Cassava Producers, Agricultural workers and Processors project in Ecuador has developed facilities to process cassava roots into 'sour starch' as a commercial enterprise, in addition to the production of 'sweet starch' for human consumption (Caballero Vera *et al.*, 1994).

Thailand is the world's major exporter of cassava and there has been good investment in mechanization for processing cassava since the 1970s. While historically the export of chips and pellets to the European Union for use in animal food has been the main focus, changes in trade agreements have resulted in diversification of cassava products and markets (Henry *et al.*, 1994). This has included the production of both native and modified starch. In the food industry, starch has been used to produce monosodium glutamate and sweeteners, with future use expected to rely more heavily on the latter, while in non-food industries, the market is in manufacture of paper, textiles and pharmaceuticals. Other Asian economies, such as that of Indonesia, also generate income from cassava starch.

As indicated above, the activity of the biosynthetic enzymes regulates the quality and quantity of the starch, with roles for mutants and gene regulation. For example, studies in potato using antisense cassava GBSS, where complete inhibition was obtained, resulted in almost completely amylose-free starch (Salehuzzaman *et al.*, 1993). Thus as applications vary according to market trends, the potential increases for research on modified starch technologies.

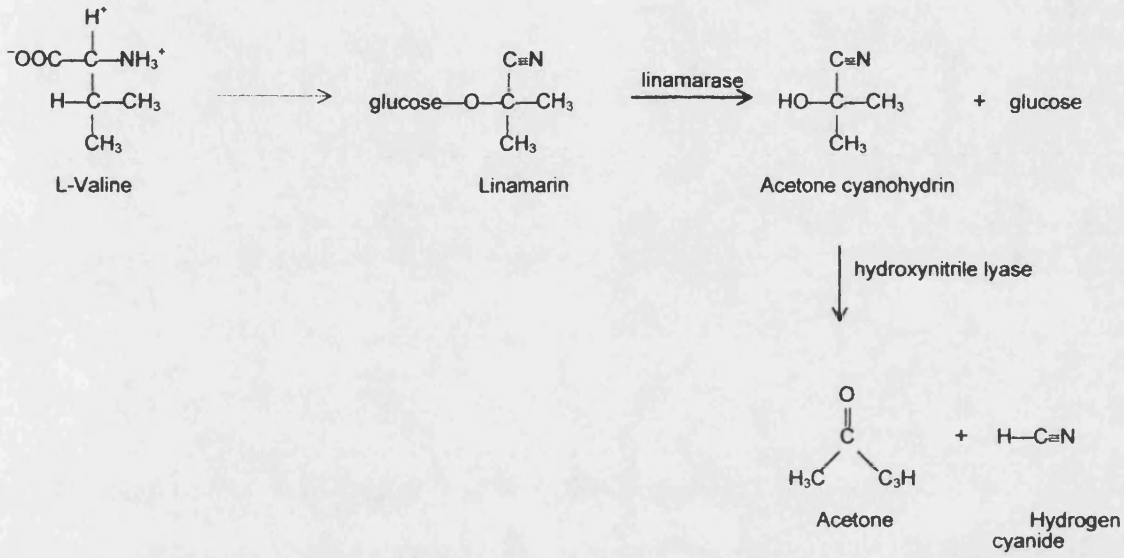
1.2.2 Cyanogenesis

One of the major problems associated with cassava is that it is highly cyanogenic, meaning that it produces hydrogen cyanide (HCN), an obvious health hazard. Cassava produces the two cyanogenic glucosides linamarin and lotaustralin from the amino acids valine and isoleucine, respectively. They are synthesized in seedling cotyledons, and it was thought that they were translocated from the leaves to all other parts of the plant for accumulation (Koch *et al.*, 1992) but some *de novo* synthesis in roots has been demonstrated (Du *et al.*, 1995). On wounding or harvesting storage roots, these glucosides undergo hydrolysis by a β-glucosidase (linamarase) to produce glucose and a hydroxynitrile (acetone cyanohydrin). The cyanohydrin is then broken down by the action of α-hydroxynitrile lyase to release the ketone (acetone) and HCN (Conn, 1980; Hughes *et al.*, 1992, 1994).

All known cultivars of cassava are cyanogenic, with linamarin accounting for around 95% of the cyanogenic glucoside level. This total amount varies between cultivars and between different roots on the same plant, and appears to be affected by the environment, particularly drought. Screening of different cultivars at the International Institute of Tropical Agriculture (IITA) in Nigeria showed a range in cyanogenic potential of 0.02–624mgkg^{-1} cyanide equivalents (Bokanga, 1995; McMahon *et al.*, 1995). Cultivars with lower cyanogenic potentials are referred to as 'sweet' or 'cool' types, while cultivars with a higher level are called 'bitter' or 'harmful' varieties.

Figure 1.5 shows a simplified pathway of HCN production.

Figure 1.5 Simplified outline of the production of HCN from the glucosides linamarin and lotaustralin in cassava (after Koch *et al.*, 1992; Hughes, 1993). Broken arrow represents multiple reaction steps via unstable intermediates ending with glucosylation to the stable glucosides. For conversion of isoleucine to lotaustralin and subsequent breakdown to release HCN, replace second methyl (CH_3) group of valine with ethyl (C_2H_5) group.



There have been many reports of a number of diseases associated with cassava consumption in Africa due to the presence of these cyanogenic glucosides, in particular, goitre, which results from poor cyanide detoxification in the thyroid, and *konzo*, an upper motor neurone disease (Cock, 1985; Tylleskär *et al.*, 1992). Therefore, lowering the level of cyanogenic compounds has been given high priority in current research. However, the fact that millions of people living in the tropics eat cassava daily without suffering such adverse effects shows that it can be consumed safely.

Processing cassava roots is a time-consuming business. They are made into different products, depending on the traditions of the area concerned and local preference. After peeling, the roots are boiled, fermented, soaked, roasted, dried, milled or a combination of these. The favoured processing method depends on a number of factors, such as the availability of firewood for roasting or water for soaking.

For example, *gari*, which is popular in Nigeria (and also eaten in Cameroon, Benin, Togo, Ghana, Liberia and Sierra Leone) is a fermented product whereby the peeled roots are grated, placed into sacks underneath heavy weights and allowed to ferment for three to four days. After sieving to remove fibres, the pulp is roasted and palm oil may be added. In Brazil, a similar method produces *farinha de mandioca*. The *gari* is then used in cooking, often soaked in water to give *eba*, a thick porridge (for a review of processing in Africa, see Hahn, 1989).

These long procedures can result in the temptation to cut corners, and it appears that this is a serious factor in cyanide intoxication. In a study comparing two areas of the Central African Republic where 'bitter' varieties of cassava were consumed daily, the area adhering strictly to traditional processing showed a lower incidence of goitre than the area where shortcuts were frequently employed (Peterson *et al.*, 1995). However, goitre is an example of an iodine-deficiency disease, therefore prevention should involve iodine supplements, in conjunction with well-processed cassava. Similarly for *konzo*, a study in rural Zaire (now Democratic Republic of Congo) showed the link between consumption of roots soaked for one day only and high dietary cyanate in children, compared with low dietary cyanate where roots were soaked for three days (Tylleskär *et al.*, 1992).

It is noticeable that such pathological conditions occur mainly in Africa, where cassava frequently forms a large part of a diet containing little protein, but are unheard of in Amazon regions of Latin America, where preference is for highly cyanogenic cassava cultivars, but the diet is supplemented with fish (Cock, 1985; Dufour, 1988). Rhodanase, which converts cyanide to thiocyanate, is a sulphur-dependent enzyme. In the absence of sulphur, it is thought that cyanate is produced, which is known to cause neurotoxicity (Tor-Agbidye *et al.*, 1999). Thus for people whose diet is deficient in sulphur-containing amino acids (often as a result of general protein deficiency) and whose reliance on cassava is heavy, the combination can produce cyanide intoxication. Unfortunately, although an excellent source of carbohydrate and vitamin C, cassava

roots contain only 1–2% protein, though the leaves contain around 7–10% (Shewry *et al.*, 1993; Bokanga, 1995). In some parts of Africa, such as the Democratic Republic of Congo, Sierra Leone and Sudan, the leaves are eaten as a vegetable, served mainly boiled as a sauce with cassava root products (Hahn, 1989).

There is a further point to be considered in the cyanide issue, which is that many farmers and consumers actually prefer the highly cyanogenic varieties. For example, studies done in both Malawi and Tanzania revealed a preference for those varieties referred to as 'harmful', which contain higher levels of the cyanogenic compounds, over the 'cool', or less cyanogenic cultivars (Chiwona-Karltun *et al.*, 1997; Kapinga *et al.*, 1997). Cassava is generally cultivated by women and if a 'harmful' variety is grown, there is less tendency for casual harvest or theft, since the root cannot be eaten without the time-consuming processing described above. This leaves the women of the family in the position of decision-making as regards the harvest of the crop and helps to ensure better food security for all family members, rather than one or two. The comment was also made that in an area where food is often scarce there is a reduced social obligation to share when visitors arrive, since soaking and pounding the roots to make 'kondowole' (the preferred product, similar to dumplings) takes several days. Reduced theft by wild animals or attack by pests, increased tuber yield, better quality of the flour obtained and tastier leaves were all mentioned as important considerations.

Thus the applications of biotechnology as regards cyanogenic compounds do not seem to favour production of an acyanogenic variety, although perhaps glucoside levels could be reduced in favoured 'bitter' cultivars while retaining desirable characteristics. The discovery of multiple UDP-glucose glucosyltransferase genes in cassava, which convert the unstable cyanohydrins to the stable linamarin and lotaustralin, may represent such a regulation point (Hughes and Hughes, 1994). It has also been observed that levels of linamarase are significantly higher in the leaves than the roots and that after cooking or processing, despite the cyanogenic potential of

leaves being 5–20 times higher than that of the roots, their cyanogenic potential was virtually eliminated (Bokanga, 1995). This implies that it is advantageous to have greater breakdown to HCN, which then evaporates during the subsequent boiling/roasting etc. Thus there is potential for increasing linamarase and hydroxynitrile lyase levels in the root, which may facilitate HCN removal.

Another area for biotechnological application is in the treatment of toxic waste, particularly where cassava processing has been mechanized. For example, waste waters from cassava starch and sago factories in India were found to contain cyanide levels from 10.4 to 27.4mgl^{-1}, the tolerable limit in effluent being 0.2mgl^{-1} (Balagopalan and Sundar, 1997). Species of *Aspergillus*, *Saccharomyces* and *Bacillus* isolated from these waters all degraded cyanide, the *Bacillus* species being most effective. In addition, anaerobic treatment of waste water, which presumably contains micro-organisms, followed by filtration through charcoal, sand and gravel gave 100% reduction in cyanide levels. Thus use of microbes, integrated with other techniques, could give low-cost methods of detoxification.

1.2.3 Post-harvest deterioration

Once harvested, cassava storage roots require processing to remove the cyanogens as described above. However, this must be done rapidly, as within 24 hours of the roots being removed from the plant they start to deteriorate. The first visible sign of this 'post-harvest deterioration' is a blue-black discolouration of the root parenchyma, known as 'vascular streaking', which later turns brown. This is followed several days later by microbial invasion and rotting (reviewed, Beeching *et al.*, 1998).

The initial part of the process is known as post-harvest physiological deterioration (PPD) and is the enzymatic, biochemical part of the root decay (Averre, 1967). The activity and levels of a number of enzymes have been shown to increase at this time. Levels of phenylalanine ammonia lyase (PAL), and the activity of the

phenylpropanoid pathway of which it is a key enzyme, also rise during PPD. This results in the production of several plant secondary metabolites, such as flavonoids and anthocyanins, lignin precursors, coumarins, phytoalexins and salicylic acid (Bennett and Wallsgrove, 1994). All of these changes have been observed associated with defense mechanisms, such as response to insects, fungi and nematodes, or wound responses, suggesting a wound-type response in cassava due to the damage sustained at harvest.

Fluorescence under UV light is associated with the vascular streaking and is thought to be due to a number of compounds. One of these is scopoletin, the level of which peaks prior to the appearance of vascular streaking and decreases with increased discolouration (Rickard and Gahan, 1983). It has anti-fungal activity and is derived via the phenylpropanoid pathway, and when applied exogenously to root tissue it rapidly causes symptoms of PPD (Wheatley and Schwabe, 1985).

Among the enzymes produced are fungal cell wall degrading enzymes, for example chitinase and β-1-3-glucanase, and a cDNA clone for the latter has been isolated from a library made from deteriorating cassava tissue (Han, Cooper and Beeching, unpublished results). In addition, peroxidases, responsible for the removal of hydrogen peroxide, increase in activity and the expression profile within the root changes. Polyphenoloxidase, which causes oxidation of polyphenols and results in browning, increases in discoloured tissue (Plumbley *et al.*, 1981).

The changes associated with the wound response are coordinated by signal transduction pathways, and although the exact mechanisms operating in the cassava storage root are not well understood, signalling molecules such as jasmonic acid, ethylene and salicylic acid are expected to be involved. Certainly it has been shown that ethylene levels increase around six hours after injury (Plumbley *et al.*, 1981), which would be consistent with a coordinating role within the overall response.

In addition to the biochemical changes, physical changes occur during PPD. Microscopy studies have shown pigmented occlusions of the xylem, probably made

from condensed tannins, which in turn are synthesized from catechins and leucoanthocyanidins in the adjacent xylem parenchyma (Rickard and Gahan, 1983). There is also reinforcement of the cell wall by synthesis of lignin, hydroxyproline-rich glycoproteins (HPRGs) and other phenolics, and a cDNA for a cassava HPRG has been isolated from the library of clones from deteriorating storage root tissue (Han, Cooper and Beeching, unpublished results). The increased peroxidase activity already mentioned has a further role here, by cross-linking HPRGs. Whereas in most plants the synthesis of these materials results in the formation of a barrier against the wound or injury, which eventually heals, in cassava this does not happen. The initial response, producing the enzymes and secondary metabolites required, is present, together with some synthesis of barrier molecules. However, complete wound healing, which normally suppresses this production via negative feedback, is absent (Beeching *et al.*, 1994). It has been speculated that this is due to the role of the storage root, which is only as a site of reserve starch, but which in other crops such as potato and sweet potato is a propagule also. Thus methods of dealing with PPD in cassava focus on its inhibition, or the enhancement of the repair mechanisms.

Different cultivars appear to show different susceptibilities to PPD, but since they are subject to different environmental and pre-harvest stresses it is difficult to quantify these levels. PPD can be markedly reduced in cassava roots by pruning the plant, which involves removing the aerial parts of the plant around two to three weeks prior to harvest. Alternatively, the harvested roots can be treated by a variety of methods. Storage at low or high temperatures inhibited PPD, e.g. refrigeration at 0–5°C or storage at 35°C, as does storage at 80–85% relative humidity (Averre, 1967). However, these procedures are not always possible. Storage in polythene bags, particularly when treated with fungicide, can inhibit deterioration in roots for two weeks or more, although it has been observed that re-injury initiates PPD, even in these treated roots (Rickard and Coursey, 1981).

The alternative to post-harvest treatment is manipulation of the levels of constitutive elements of the PPD pathways, such as PAL or scopoletin. However, the problem here is that by altering the early part of the defense pathway, microbial infection could be accelerated, so perhaps targetting the lignification reactions by means of peroxidases and polyphenoloxidases would be more appropriate. Increased lignification and suberization would provide a barrier against attack by microbes and also reactive oxygen species. These latter species are potentially harmful to plants, yet are themselves involved in defense, paradoxically causing cross-linking of HPRGs and lignification, and stimulating phytoalexin synthesis. The antioxidants which detoxify reactive oxygen species involve compounds that are easily oxidized, such as phenolic compounds, coloured compounds which absorb potentially damaging wavelengths, such as carotenoids, and chelating agents, such as citric acid (Larson, 1995). Some cassava roots contain more β-carotene than others, giving them a yellow colour, and there are reports that these are less prone to PPD. IITA has developed a number of cultivars with high β-carotene content, which are used to make 'yellow *gari*', normally produced by adding (expensive) palm oil to *gari* made from white roots (IITA, 1990). Additionally, these improved varieties were reported to have good resistance to pests and diseases. However, little is understood about the underlying control of this complex set of reactions in cassava, and more information is needed before biotechnological improvements can be effectively targetted to control PPD without compromising the resilience of the plant.

1.2.4 Pests and diseases

Cassava is prone to various pests and diseases, all of which can cause severe losses of the crop. Of the pests, most devastation is caused by the cassava mealybug (*Phenacoccus manihoti* Mat. Ferr.) and the green spider mite (*Mononychellus tanajoa* Bondar). Both of these are relatively recent introductions to Africa, with the first reports

of their damage during the 1970s (Herren, 1994). In their native Latin America they have co-evolved with natural control mechanisms, which are absent in Africa. Hence both have been subjected to concerted biological control programmes using imported predators; the endophagous parasitoid *Epidinocarsis lopezi* (De Santis) against mealybug, and phytoseiids and pathogenic fungi in the case of the green mite (Neuenschwander, 1994; Yaninek, 1994). This has been particularly successful in treating mealybugs.

African cassava mosaic disease (ACMD) is a viral disease of cassava, spread by the whitefly *Bemisia tabaci.* Causing leaf discolourations and lesions, stunting of the crop and loss of yield, it results in massive crop failure throughout Africa and there have been several recent epidemics. Serological testing has shown several variants of the geminivirus causing the disease, with clear regional distinctions, leading to the classification into African cassava mosaic virus (ACMV), East African cassava mosaic virus (EACMV) and Indian cassava mosaic virus (ICMV). These three strains have been assumed to correlate with the three points of entry of cassava to coastal West Africa, Madagascar and East Africa, and Sri Lanka and India, respectively (reviewed Thresh *et al.*, 1994).

The most recent epidemic of 1996 swept through central Africa, particularly Uganda, with the 'front' travelling at 20–30 miles per year. This was accompanied by massive increases in the numbers of whitefly observed. Studies amplifying part of the virus using the polymerase chain reaction (PCR) were carried out 'ahead of', 'at' and 'behind' the front (Deng *et al.*, 1997). 'Ahead of' the front were healthy looking cassava plants, with some mild symptoms of mosaic disease, a low whitefly population and normal cassava production. 'At' the front, the whitefly numbers increased, as did the severity of the symptoms, while 'behind' the front tuberous root production was extremely limited. The DNA sequence of the virus at these three locations had a number of differences from ACMV but was very similar to EACMV. However, the coat protein showed similarity with EACMV at the N-terminus, but near identity to ACMV at

the C-terminus. These results indicate the presence of a new variant of these two geminiviruses. PCR-based typing techniques could be extremely useful in identifying the current strain or hybrid of a particular epidemic, which may then allow the selection of resistant cultivars, as has been done by Zhou *et al.* (1997).

Breeding programmes for disease resistance at institutes such as CIAT in Colombia, IITA in Nigeria and the Namulonge Agricultural and Animal Research Institute (NAARI) in Uganda identify resistant clones in field trials with release of cuttings to local farmers (Otim-Nape *et al.*, 1994). Three varieties developed at NAARI from tissue culture supplied by IITA have shown some resistance to the new-variant epidemic of ACMV and have proved to be acceptable alternatives to local cultivars. Nase I, Nase 2 and Migyera were adopted by local farmers in Uganda, where cassava is the second most important crop after bananas (Ssemakula *et al.*, 1997).

ACMV does not become fully systemic in certain 'resistant' cultivars, which means that the spread of disease from cuttings is slow, with some cuttings being virus-free (termed 'reversion'). By using reversion and cutting selection, disease incidence equilibrium values can be reached which are less than the 100% achieved when they are not practiced, (Fargette *et al.*, 1994), underlining the importance of good phytosanitation practice. In addition, farmers have been willing to change cultivars to help limit the disease, select healthy planting material and 'rogue' plants (remove infected individuals from the field).

Cassava bacterial blight, caused by *Xanthomonas axonopodis* pv. *manihotis* (formerly known as *Xanthomonas campestris* pv. *manihotis*), occurs worldwide and symptoms include angular leaf lesions, wilting, stem exudates and cankers, and dieback, resulting in crop losses of up to 100% (Boher and Verdier, 1994). Originating in South America, it was first observed in Africa during the 1970s. The bacterium encounters the plant during the rainy season, transferred by rainwater or insects. After landing on the leaves, it multiplies and when sufficient inoculum is present infects the tissue. This is via the lower stomata which, unlike the upper stomata, are not occluded

by wax and are concentrated along the mid-rib and major veins (Cooper *et al.*, 1997). The disease quickly becomes systemic, which causes the problem of transmission by stem cuttings.

As with ACMV, treatment involves a combination of phytosanitary approaches and breeding for resistance, using *M. esculenta* and a wild relative, *M. glaziovii* (Hahn *et al.*, 1980). Molecular analyses have been used in typing strains of the pathogen, using restriction fragment length polymorphisms (RFLP) of the *pth*B gene, identified as being involved with pathogenicity (Restrepo *et al.*, 1997). Pathogenic strains of *Xanthomonas spp.* have also been typed by monoclonal antibodies and PCR, and research is underway to provide molecular markers on the cassava genetic map to help with breeding disease-resistant cultivars (reviewed; Verdier *et al.*, 1997).

1.2.5 Availability of improved planting material

As mentioned above during the discussions of post-harvest deterioration (**1.2.3**) and diseases (**1.2.4**) the roles of the various research institutes in providing improved planting material is extremely important. Farmers appear willing to try new cultivars, provided certain characteristics are maintained, mainly to do with taste and texture. Particularly in response to epidemics, there is a need for rapid dissemination of material to farmers. Since cassava is mainly propagated by stem cuttings, the generation of large numbers of plants is limited by the number of resistant individuals developed initially, which is a slow process. However, micropropagation provides an alternative and is a good way of producing planting material free from disease.

Micropropagation systems are well established in many countries and have been used for *in vitro* multiplication. The use of meristem cultures from axillary buds has increased potential for the production of new plants. Culture on media containing cytokinins, with or without the addition of auxins, allows proliferation of multiple shoots,

with up to 25 shoots obtained in the best responding cultivars (review, Puonti-Kaerlas, 1998).

1.2.6 Conservation of cassava diversity

Among the first observations on the origins of agriculture and the diversity of crops were those made by the Soviet scientist Nikolai Vavilov in the early part of the 20th century. During his travels across the globe collecting seed he observed that diversity was not evenly distributed. Certain areas showed large numbers of endemic varieties, such as the hundreds of ancient wheat varieties found on a plateau in Ethiopia, whereas other areas showed little variation. Vavilov reasoned that a crop grown in an area for a long period of time had opportunities to show diversity, with cultivars being selected on the basis of differing characteristics for different purposes. He therefore concluded that centres of diversity represented areas of origin, which he noticed overlapped for many crops. Eight centres of origin for agriculture were proposed, all in what is now termed the 'developing world' and often in mountainous regions that provide their own natural barriers (reviewed in Fowler and Mooney, 1990). These ideas have subsequently been revised, with not all crops having their origin in the area of most diversity, but the original concept is still useful with increased variety being associated with ancient cultivation. For example, wheat is not thought to have originated in Ethiopia, despite the diversity found there. Certain centres of origin for crops in general are accepted – the eastern Mediterranean, northern China and central America – with so-called 'secondary centres' of diversity being established for some crops as they spread out from the original areas.

Two centres of biodiversity have been proposed for cassava, the major one in northeastern Brazil and a secondary one in Mexico. There are 98 reported species of the genus *Manihot* and cassava is found in all areas where these are present, though as a crop it is grown within the bands of 30°N and 30°S of the equator (Rogers and

Fleming, 1973). After the colonization of Brazil by the Portuguese, cassava is proposed to have travelled to their colonies in West Africa around 400 years ago and to Goa in the 18th century, with a further point of entry to Africa on the east coast. The spread to Indonesia and the Philippines is also thought to have happened at this time, from Mexico, with subsequent dispersal throughout the tropics (Cock, 1985).

The domestication of cassava is unclear but it is thought to have arisen from wild relatives that produce tuberous roots. It has a chromosome number of 36 in the diploid state, which suggested its origins as being from two closely related taxa. Studies based on chloroplast DNA and the intergenic spacer region of nuclear ribosomal RNA suggested *Manihot esculenta* ssp. *flabellifolia* as a progenitor, followed by intensive selection (Fregene *et al.*, 1994). This species grows in the Brazilian forest, along the southern and eastern borders of the Amazon basin, as a clambouring shrub. Recent studies using the single copy nuclear gene glyceraldehyde 3-phosphate dehydrogenase (*G3pdh*) with cassava, *M. esculenta* ssp. *flabellifolia* and *M. pruinosa*, a closely related species, suggest that domestication occurred from ssp. *flabellifolia* along the southern Amazon basin (Olsen and Schaal, 1999).

It is thought that the lack of genetic diversity in Africa has contributed to the level of devastation and crop losses caused by outbreaks of pests and diseases (Dahniya, 1994), and the introduction of other varieties from Latin America could prove useful. However, maintenance of local germplasm within collections organized by the research institutes, such as the programmes at CIAT and IITA are extremely important, particularly when non-local cultivars, which may be desirable from disease-resistant or cooking quality perspectives, are grown in preference. The concern is that cultivars may be rapidly and permanently lost, since seed is rarely collected for propagation in the field.

Currently CIAT holds almost 6000 accessions, the largest *in vitro* cassava collection in the world. At present these are maintained under slow growing conditions, but the most stable form of storage is cryopreservation. Seeds, somatic embryos,

meristems and shoot tips can all be preserved in this manner, but there are difficulties associated with the regeneration of certain cultivars (review, Puonti-Kaerlas, 1998). The most common method of regeration for cassava is somatic embryogenesis, first described for cotyledons and embryogenic axes by Stamp and Henshaw (1982). The frequency of regeneration of mature somatic embryos is low and a number of improvements have been made. The usual procedure now involves primary somatic embryogenesis followed by a cyclic embryogenesis step in liquid culture, resulting in faster, increased production than regeneration on solid media (Stamp and Henshaw, 1987; Raemakers *et al.*, 1993). A major development has been the production of an alternative embryogenic culture, where the tissue proliferates in liquid culture to produce a friable embryogenic callus (FEC), consisting of thousands of embryogenic units (Taylor *et al.*, 1996). This gives a high quality suspension culture which allows regeneration to plantlets, with the potential to produce large numbers of plants from the original.

1.2.7 Genetic modification of cassava

With the advent of stable transformation procedures for cassava, the potential for improving cassava by the manipulation of genes of interest has become a real possibility. Both *Agrobacterium tumefaciens*-mediated transfer (Li *et al.*, 1996) and the insertion of 'naked' DNA via biolistic transfer (Schöpke *et al.*, 1996) have been successful. The best tissue for the latter procedure is the FEC described above, but at present there is a low regeneration rate. However, as conditions are optimised this is likely to improve.

A genetic linkage map for cassava has been published, using a variety of molecular markers. Restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPDs), microsatellites and isoenzymes were used on the F_1 cross between cultivars from Nigeria and Colombia (Fregene *et al.*, 1997). It is

difficult to breed cassava true to type from seed as it is an allo-tetraploid and out-breeding, and hence genotypes are likely to be heterozygous (Cock, 1983). Thus a linkage map is important for biotechnological applications.

It is clear that, as for all major crops, as many strategies as possible should be used to combat the numerous problems encountered by farmers and consumers. All the above limitations of cassava could benefit from a molecular biological approach in conjunction with the more traditional methods. The introduction of foreign storage protein genes, for example, could be useful in overcoming dietary problems associated with heavy reliance on cassava as a staple, but more immediate remedies such as iodine dietary supplements and adhering to strict processing procedures would alleviate the major health risks.

Whether aiming to alter the regulation of those genes already present or to insert others from different species, knowledge of constitutive promoters and control elements will be extremely important. A cDNA clone of linamarase from cassava has been used to obtain the genomic sequence and the promoter region is being studied and used linked to β-glucuronidase (GUS) in particle-gun bombardment assays to transform root and leaf tissue (Liddle *et al.*, 1997). It appears to express GUS only in the roots, although mature tubers were not transformed. As with other crops, tissue- or developmental-specific promoters will be particularly useful.

1.3 Overview of this project

As mentioned earlier in the discussion of cyanogenesis (**1.2.2**), cassava storage roots are poor in protein, with low levels of the sulphur-containing amino acids methionine and cysteine. There is therefore the potential to increase both the protein quantity and quality of the storage root. At the same time, promoters for cassava genes would be useful in genetic modification and avoid the reliance on viral promoters. These often require co-transformation of selectable marker genes such as

those for antibiotic resistance, with the accompanying risk of the spread of antibiotic resistance through crop and weed populations. Therefore, identification of a promoter for a storage root protein was considered a logical choice.

1.3.1 Storage proteins found in other roots and tubers

Storage proteins are synthesized by most plants to act as a repository for nitrogen, sulphur and carbon. They are mainly found in those tissues or organs involved in reproduction, propagation and dispersal, although they are also made when nutrients are plentiful for times of future hardship (for a review of storage proteins, see Shewry, 1995). While they have been more extensively studied in seeds, they are also under investigation in crops producing tubers, such as potato (*Solanum tuberosum* L.), sweet potato (*Ipomoea batatas* (L.), Lam.), yam (*Dioscorea* spp., of which *D. rotundata* and *D. cayenensis* Lam. are the most important) and taro (*Colocasia esculenta* L.).

Potato tubers contain a family of glycoproteins of approximately 40kDa and known as patatin, which constitute around 40% of the total soluble protein (Park, 1983). They divide into class I and class II patatin genes which are 98% homologous but differentially expressed (Mignery *et al.*, 1984). There are 10–15 copies per haploid genome, not all of which are active genes, with class I being predominantly expressed in the tuber, while class II are found in roots and tubers but at much lower levels than class I transcripts (Pikaard *et al.*, 1986, 1987).

Sporamins A and B account for over 80% of the soluble protein of sweet potato tuberous roots and have a molecular weight of around 20kDa, but unlike patatin are not glycosylated (Hattori *et al.*, 1985; Maeshima *et al.*, 1985). Again, they are encoded by a multigene family of about 10 genes per haploid genome, with over 80% homology in the coding regions of the two sub-families, but highly diverged 5'-upstream regions (Hattori *et al.*, 1989; Ishiguro and Nakamura, 1992).

Around 85% of the total protein in yam tubers is comprised of the dioscorins, which have a molecular weight of 31kDa and divide into two classes, showing around 84% similarity in the coding regions (Harvey and Boulter, 1983; Conlan *et al.*, 1995). Taro has two globulin families, G1 of 12.5kDa and G2 with two subunits of 22kDa and 24kDa, which account for around 80% of the total soluble tuber protein and are proposed to have a storage function (Monte-Neshich *et al.*, 1995).

1.3.2 Expression patterns

These various storage proteins are located in the tuber vacuoles, and under normal conditions there is little, if any, expression outside of the tuber, the exception being class II patatin transcripts (Hattori *et al.*, 1988; Sonnewald *et al.*, 1989; Prat *et al.*, 1990; Conlan *et al.*, 1995; Monte-Neshich *et al.*, 1995). They do not, however, show homology between species, which probably reflects the different origins of the tubers themselves and therefore the potential origins of the proteins. Whereas potato tubers develop from a lateral shoot, or stolon, from the underground stem, sweet potato tubers are swollen roots, which is also the case for cassava. Yam tubers develop from the hypocotyl and taro has corms, which are enlarged stems, but differ from the potato tuber since they undergo radial expansion at the same time as tuber formation (summarized in Shewry, 1995; Monte-Neshich *et al.*, 1995.)

1.3.3 Roles beyond the storage of nitrogen and sulphur

1.3.3.1 Patatin

Further to the differences in glycosylation and molecular weights mentioned above, patatin has lipid-acyl hydrolase activity (Racusen, 1984; Rosahl *et al.*, 1987) and it has since been demonstrated that patatin is a cytosolic phospholipase A(2) (Senda *et al.*, 1996). These results may indicate a possible role in defence against

microbial attack (Prat *et al.*, 1990) which is given further support by the characterization of a latex allergen Hev b7 from rubber (*Hevea brasiliensis*) with homology to patatin (Beezhold *et al.*, 1997, Kostyal *et al.*, 1998), since other defence-related proteins in latex are also allergens.

1.3.3.2 Sporamin

Sporamin also appears to have a role when the tubers are damaged, either mechanically or by pest or microbe attack, since it is homologous to Kunitz-type proteinase inhibitors, the poplar *Win3* gene product and cathepsin D inhibitor from potato, these latter two being wound-inducible (Bradshaw et al., 1989; Hattori *et al.*, 1989; Suh *et al.*, 1991). It was also found to be induced in cut sweet potato leaves when these had been dipped in polygalacturonic acid or chitosan (Ohto *et al.*,1992), known inducers of wound-related proteinases (Ryan, 1981). Functional trypsin inhibitory activity has recently been demonstrated for sporamin and its two precursors, preprosporamin and prosporamin (Yeh *et al.*, 1997). Wounding resulted in the accumulation of sporamin transcripts in other non-damaged parts of the plant, with 2–3 times the level in unwounded leaves than wounded leaves. This correlates well with patterns obtained for proteinase inhibitor gene expression in both potato and tomato, implying strongly that sporamin is indeed a protein involved in wound response.

1.3.3.3 Dioscorin

In yam, similarity at the DNA level has been found between dioscorin and animal carbonic anhydrase, the enzyme responsible for catalysing the hydration of carbon dioxide, although at the amino acid level mutations have occurred in dioscorin within the area corresponding to the active site (Conlan *et al.*, 1998).

1.3.4 Cassava storage proteins

Unlike these other tuber and root crops, cassava does not appear to have storage proteins (Shewry *et al.*, 1993). The fact that the storage root of cassava is not a propagule may well explain this, as in contrast to the above examples, there would be no requirement to ensure a ready source of protein for growth and cellular division. Rather the storage root in cassava is precisely that – a carbohydrate sink in times of excess to be used by the mature plant in times of stress – and has a mere 1–2% protein contributing to the overall dry weight (Yeoh and Chew, 1977).

1.3.5 Investigation of a small cassava protein located in the storage root

The aim of this project was originally to obtain genetic information regarding a small (25kDa) protein from the tuber parenchyma. This protein appeared to be up-regulated by nitrogen, and further studies on the effect on the storage root with different nutrient solutions were carried out (Chapter 3). An antibody raised to the protein had been purified, and it was hoped this could be used to identify a clone from a cDNA expression library, which in turn would allow the isolation of the genomic clone. This would then give the opportunity to analyse the promoter region of a gene expressed in the storage root. However, a low insert frequency within the cDNA library to be used meant the construction of a new storage root cDNA library, discussed in Chapter 4. Once this had been made, serious problems were found with the antibody to be used and this strategy was therefore abandoned.

1.3.6 Alternative strategy for the identification of clones expressed in the storage root

An alternative screening method was employed, using PCR to detect recombinant molecules. Although this approach picked out cDNA clones at random the advantage was that they were known to be expressed in the storage root.

Additionally there were very few accessions from cassava on the database, with those studied so far from particular areas such as the starch biosynthetic pathway (Salehuzzaman *et al.*, 1992, 1993; Munyikwa *et al.*, 1994, 1997) and cyanogenesis (Hughes and Hughes, 1994; Hughes *et al.*, 1992, 1994). Thus there was a good chance that clones identified in this manner would lead to some interesting, and possibly novel, genes from the storage root.

The clones identified from the cDNA library are discussed in Chapters 5, 6 and 7, the latter chapter dealing with most interesting clone which was chosen for analysis of the gene.

CHAPTER 2: GENERAL MATERIALS AND METHODS

All chemicals used were of Analar grade and supplied from Sigma (USA) or BDH (UK) unless otherwise indicated. Media for bacterial culture was obtained from Lab M (UK). Recipes for commonly used buffers and culture media are given in Appendix 1 and information regarding the various cultivars used in the project is given in Appendix 2.

2.1 Cassava growth conditions

2.1.1 Growth period

For the nutrient studies discussed in Chapter 3, 40 cuttings of the cassava cultivar MCol 1684 were taken and grown up in size LT pots for two weeks. After this time, 24 were taken to be re-potted into size 9 pots in a mix of 50:50 perlite:sand and grown for a year in a heated greenhouse, with a constant temperature of 25°C and 70% relative humidity, and a 12h light /12h dark cycle.

Tissue for all other experiments was obtained from the appropriate cultivar, using plants propagated and grown in a similar manner to that described above.

2.1.2 Nutrient regimes

Six plants were used as replicates in each nutrient group for the studies in Chapter 3. All plants were fed on Monday, Wednesday and Friday, each group of six replicates with one of the four nutrient solutions detailed in Table 2.1.

Table 2.1 Composition of the low and high nitrogen nutrient solutions (gl^{-1})

	LN–S*	HN–S†
KCl	0.522	—
$CaCl_2.6H_2O$	0.438	—
$Ca(NO_3).4H_2O$	0.472	1.18
CH_2PO_4	0.136	—
KH_2PO_4	—	0.136
KNO_3	—	0.708
$MgCl_2.6H_2O$	—	0.096
$Mg(NO_3)_2.6H_2O$	0.410	0.439
NaCl	0.0059	0.059
NH_4Cl	—	0.107
EDTA FeNa	0.018	0.018
Trace Elements	0.5ml	0.5ml

*LN+S: As LN–S, plus $MgSO_4.7H_2O$ 37.39g
†HN+S: As HN–S, plus $MgCl_2.6H_2O$ + $MgSO_4.7H_2O$ 37.40g

Trace Elements (gl^{-1}):

Zinc acetate	0.42
H_3BO_3	3.045
$Cu(NO_3)_2.3H_2O$	0.24
Ammonium molybdate	2.025
EDTA FeNa	36.03
$Mn(NO_3)_2.6H_2O$	5.56

Alternative days and at weekends they were watered with de-ionized water to prevent the build up of salts and to flush the root systems. The pots were filled sufficiently to obtain run-off from the base.

2.1.3 Pesticides

Biological control agents and chemical pesticides were both used on the growing cassava plants. Mycotal (Koppert, The Netherlands), containing the entomopathogenic fungus *Verticillium lecanii*, was used against aphids, at a concentration of $1.5gl^{-1}$. This was applied weekly during June and July, while the new cuttings were growing up, and fortnightly from October to April. Phytoseiulius (Avon Crops, UK), a red spider predator, was also used once. The chemical pesticide Torque (fenbutatin acid; ICI, UK) was used regularly; once a month for the first 7 months, then weekly during February and April. Temik (aldicarb; Union Carbide, UK) was used once and Nemolt (teflubenzuron; Cyanamid, UK) twice.

2.1.4 Harvest of storage root and leaf tissue

For the nutrient studies, at the end of a year's growth the aerial parts were removed, the plants uprooted and the number of roots on each replicate noted. The appearance of the plants and their roots was noted, and the roots counted and measured. Roots were then carefully removed individually, washed in distilled water and quickly weighed before being frozen in liquid nitrogen for storage at –70°C.

Storage root tissue for mRNA preparation was harvested after approximately one year's growth, as required, i.e. it was possible to harvest part of the storage root and leave the remainder growing on the plant. Generally entire roots were harvested, with the excess tissue stored at –70°C. Leaf tissue was harvested earlier from cultivars, choosing young, newly opened leaves to avoid having to remove excess photosynthate during nucleic acid preparations.

2.2 cDNA library construction

2.2.1 mRNA preparation

2.2.1.1 Pre-treatment of equipment

Prior to starting RNA extraction, most of the glassware to be used was acid-washed for 24h in 5M H_2SO_4. After being thoroughly rinsed, and together with the grater, spatulas, pestle and mortar, this glassware was baked in aluminium foil, for a minimum of 2h at 200°C to destroy any ribonucleases. Plasticware, together with glassware such as volumetric flasks which could not be baked at high temperatures, were soaked in 1M NaOH overnight, then rinsed with sterile MilliQ water (SMQ; Millipore, France) which had been autoclaved at 15psi for 30min. It was decided not to use diethylpyrocarbonate (DEPC) to treat the items to be used, at least initially, since

this can cause RNA degradation if not completely removed and is also hazardous to health.

All the solutions required used chemicals reserved solely for RNA work, to avoid possible contamination, and were made up 'blind', using pre-determined amounts of acid or alkali to obtain the desired pH. Autoclaved SMQ was used in all solutions.

2.2.1.2 Total RNA extraction

This protocol was adapted from that in the Plant Molecular Cloning Manual (Pawlowski *et al.*, 1994).

Approximately 5g of young leaves, or 10g of storage root parenchyma, were harvested from various plants of the CMC 40 cultivar. Leaves were frozen immediately in liquid nitrogen before being ground to a fine powder in a pre-cooled pestle and mortar. Storage root sections were peeled with a sterile, pre-cooled knife, grated using a sterile grater directly into liquid nitrogen in a pestle and mortar, similarly pre-cooled. Remaining portions of root were stored at –20°C until required. $4ml g^{-1}$ of RNA extraction buffer (100mM LiCl, 1% SDS [v/v from a 10% stock], 100mM Tris-HCl, 10mM EDTA, pH9) were warmed to 60°C and added to the leaf powder. $2ml g^{-1}$ each of phenol and chloroform were then added, the mixture vortexed for 2min after each addition. During the first root extraction, $8ml g^{-1}$ of extraction buffer and $4ml g^{-1}$ each of phenol and chloroform were used. Only three phenol:chloroform extractions were performed since there was no visible protein at the organic:aqueous interface. Since this procedure resulted in large volumes for the overnight precipitation step, subsequent extractions were performed using the same volumes as for the leaf RNA extraction.

The extraction was centrifuged at 12700rpm for 30min at 4°C (Sorvall RC5B centrifuge, SS34 rotor) and the aqueous phase removed to a fresh tube. An equal volume of phenol:chloroform was added, vortexed for 2min and centrifuged as before.

This step was repeated three times and the resulting aqueous phase extracted once with an equal volume of chloroform alone. After centrifugation as above, the RNA was precipitated overnight at 4°C by the addition of lithium chloride to a 2M final concentration. The RNA was pelleted by centrifugation at 9200rpm for 10min as above, then washed twice with 70% ethanol before being resuspended in 300µl 0.3M sodium acetate, pH 5.2. This was extracted with an equal volume of chloroform then microfuged at 13461*g* at room temperature for 5min. The aqueous phase was removed and the RNA precipitated in 2.5 volumes of 100% ethanol, incubated on ice for 15min. This was centrifuged, again at 13461*g* for 5min at room temperature and finally washed in 70% ethanol. The resulting pellet was dried under vacuum, resuspended in 100µl SMQ and stored at –70°C.

2.2.1.3 RNA gel electrophoresis

The quality of the RNA preparations were checked by gel electrophoresis on a 1.2% (w/v) agarose:2.2M formaldehyde gel, cast and run in a flow hood.

Samples were prepared in 20µl total volume, containing 0.5× running buffer (1× running buffer: 20mM 3-(N-morpholino) propanesulphonic acid [MOPS] pH7.0; 8mM sodium acetate; 1mM EDTA), 6.5% (v/v) formaldehyde, 50% (v/v) formamide, and incubated at 65°C for 15min. They were chilled on ice, then centrifuged for 5s. 2µl of loading buffer (50% (v/v) glycerol, 1mM EDTA, 0.25% (w/v) bromophenol blue, 0.25% (w/v) xylene cyanol, pH8) were added and also 1µl of a 1mgml^{-1} stock solution of ethidium bromide. The samples were immediately loaded onto a gel pre-run at 3Vcm^{-1} for 5min, then run overnight in 1× running buffer at 1Vcm^{-1}. The gel was viewed under UV light at 320nm, using a UVP White/UV Transilluminator (UVP Life Sciences, Cambridge), and photographed.

2.2.1.4 Purification of total RNA preparation

Tip-100 filters (Qiagen, UK) were used to purify the crude RNA preparations. Each tip is suitable for up to 200μg RNA. The total RNA pellet was dissolved in 8ml R5 buffer (20mM Tris-HCl, 1mM EDTA pH8) and 2ml R6 buffer (2M NaCl, 250mM MOPS pH 7). The Tip-100 was equilibrated with 3ml QAT buffer (400mM NaCl, 50mM MOPS, 15% ethanol, 0.15% Triton-X pH7) and the sample applied to the column. 30ml QA buffer (QAT buffer without Triton-X) were used to wash the sample before eluting it with 10ml QRU buffer (1.2M NaCl, 67mM MOPS, 20% [v/v] ethanol, 6M urea pH6.7). The RNA was precipitated with an equal volume of isopropanol and centrifuged at 9638*g* for 30min at 4°C. The resulting pellet was washed in 70% ethanol, air-dried and then redissolved in SMQ.

All buffers were filter sterilized since they contain MOPS which breaks down on autoclaving.

2.2.1.5 Isolation of poly (A^+) RNA

Poly (A^+) mRNA was isolated by means of a 1ml oligo (dT) cellulose column (Pharmacia, USA), on which up to 10mg total RNA can be bound. It was necessary to reduce the column volumes for these isolations, as approximately 220μg leaf and 450μg storage root RNA were used. First the resin was suspended in 0.1M NaOH and the column poured in a sterile glass pasteur pipette, stoppered with siliconized glass wool. This was then washed with 3 column volumes of SMQ, before being equilibrated to pH<8 using 1× loading buffer (20mM Tris-HCl pH7.6, 0.5M NaCl, 1mM EDTA pH8, 0.1% sarkosyl). The RNA sample was heated to 65°C for 5min, cooled rapidly to room temperature and an equal volume of 2× loading buffer added. This was applied to the column and the eluate collected. A single column volume of 1× loading buffer was used to wash the column and again, the eluate collected before being heated to 65°C for 5min and reapplied to the column. This final eluate was collected.

The column was washed with 5–10 volumes of 1× loading buffer and 0.5ml fractions collected for spectrophotometric analysis at A_{260}. Initially, the absorbance reading is very high, as the non-polyadenylated RNA passes through the column, while the later fractions should have little or no absorbance.

The poly (A^+) RNA was eluted with 2–3 column volumes of elution buffer (10mM Tris-HCl pH7.6, 1mM EDTA, 0.05% SDS [v/v from a 10% stock]) and fractions collected of 0.3–0.5 column volume. After measuring the A_{260} in cuvettes previously washed in a 1:1 mix of concentrated HCl:methanol for a minimum of 1h, the fractions containing RNA were pooled. This step was omitted in later purifications, since the RNA levels in the preparations were approaching the level of sensitivity of the spectrophotometer. As material obtained after the first round of chromatography contains approximately equal volumes of poly- and non-polyadenylated RNA, the sample was heated to 65°C, rapidly cooled, the concentration of NaCl adjusted to 0.5M and the eluted sample reapplied to the column for a second round of purification as above.

A tenth volume of 3M sodium acetate (pH5.2) and 2.5 volumes of ice-cold ethanol were added to precipitate the RNA, which was left on ice for a minimum of 30min. The sample was then centrifuged at 9638*g* for 15min at 4°C and the pellet washed in 70% ethanol, air-dried and redissolved in a small volume of SMQ. This was stored at –70°C under 3 volumes of ethanol. When required, the RNA was recovered by the addition of a tenth volume of sodium acetate (pH5.2) and centrifugation at 9638*g* for 5 min at 4°C.

2.2.1.6 RNA quantification

RNA quantity was calculated using the results from the spectrophotometric absorbance readings at 260nm, according to the following equation:

$$A_{260} = 1 \equiv 40\mu g ml^{-1} \text{ RNA}$$

The purity was assessed by calculating the A_{260} : A_{280} ratio, where a result of 1.8–2.0 indicates no protein contamination. These calculations were performed at each stage in the poly (A^+) extraction.

2.2.2 cDNA synthesis

The cDNA synthesis module from Amersham (UK) was used. First strand cDNA synthesis used 1µg of the substrate RNA, thawed on ice, with the following reaction components added from the kit: first strand buffer (50mM Tris-HCl pH8.3, 50mM KCl, 10mM $MgCl_2$), 4mM sodium pyrophosphate, 1 unit human placental ribonuclease inhibitor, 1mM each dATP, dGTP, dTTP, 0.5mM dCTP, 7µM random hexanucleotide primers, 5µCi α–^{32}P–dCTP. These were mixed gently and centrifuged briefly before adding 20 units of Avian Myeloblastosis Virus reverse transcriptase and incubating at 42°C for 1h. The first-strand synthesis reaction was then placed on ice and a 1µl sample removed to estimate incorporation of the radionucleotide.

Second strand synthesis was carried out by adding the following components in the order given to the first-strand reaction: second strand buffer (16mM Tris-HCl pH7.5, 80mM KCl, 4mM $MgCl_2$, 0.05mgml^{-1} bovine serum albumin [BSA]), 50µCi α–^{32}P–dCTP, 0.8 units *E.Coli* RNase H, 23 units *E.Coli* DNA polymerase I and sterile water (all supplied in the kit) to a final reaction volume of 100µl. These were mixed gently and incubated for 1h at 12°C, 1h at 22°C and 10min at 72°C. After microfuging briefly, the reaction was placed on ice and 2 units T4 DNA polymerase added. This

was incubated for 10 min at 37°C, then the reaction was stopped by adding 0.25M EDTA to a final concentration of 0.01M. A 1µl sample was removed for analysis by Cerenkov counting. The amount of radioactivity used in the construction of the cassava leaf cDNA library was that recommended by the Amersham protocol and outlined above. For the storage root cDNA library, one tenth the recommended amount was added, i.e. 0.5µCi and 5µCi for the first- and second-strand synthesis reactions respectively.

The double-stranded cDNA was purified by two phenol:chloroform extractions and a single extraction using chloroform alone, then precipitated at –20°C overnight. After warming to room temperature, the reaction was microfuged at 9638*g* for 10 min, and the pellet washed with 50µl of 2M ammonium acetate and 100µl ethanol before re-centrifugation. A further washing with 200µl ethanol was performed and the final pellet dried for 15 min at 37°C and redissolved in 10µl TE buffer (10mM Tris-HCl pH8, 1mM EDTA).

The two 1µl samples removed from each of the first and second strand synthesis reactions were diluted to 5µl with sterile water. 2µl of these samples were then spotted onto DE 81 filters (Whatman, UK). Two filters, A and B, were used for each reaction. After air drying, the two B filters were washed 6×5 min in 0.5M Na_2HPO_4, followed by 2×1min washes in sterile water then 2×1min washes in 95% ethanol. The filters were allowed to dry thoroughly, then all filters were counted by Cerenkov counting. Control filters with non-radioactive spots were also processed to monitor the washing procedure and a final blank with no filter used to estimate background radiation.

2.2.3 Addition of adaptors to cDNA

After cDNA synthesis, *Eco*RI adaptors were ligated on to the ends of each cDNA molecule to facilitate cloning into the bacteriophage λgt11 which has a unique

*Eco*RI cloning site (Amersham cDNA rapid adaptor ligation, rapid cloning and λ DNA *in vitro* packaging modules). For every 1µg cDNA, 250pmol adaptors and 12.5 units of enzyme enhancer were used. Controls using cDNA were not performed for the remaining steps of the library construction, since this requires the removal of 30% of the original DNA sample, but the blunt-ended control DNA supplied went through the same procedures as the cassava cDNA. 16 units of T4 DNA ligase were added and the reaction incubated for 30min at 16°C, then stopped by the addition of EDTA to a final concentration of 0.025M (pH8 from a 0.25M stock). STE buffer (10mM Tris-HCl pH8, 1mM EDTA, 100mM NaCl) was added to create a final sample volume of 100µl.

After storage at –20°C overnight, the 'adapted' DNA was passed through spun columns to purify and size fractionate the cDNA. The columns were equilibrated with STE buffer then centrifuged at 1500*g* for 2min. The cDNA sample was then carefully applied to the centre of the resin and centrifuged as before.

2.2.4 Kinasing reaction

Since the ends of the adaptors are unphosphorylated (to minimise self-ligation) it is necessary to perform a kinasing reaction. 32 units of T4 polynucleotide kinase were added to the eluted purified DNA in a final buffered reaction volume of 200µl. This was incubated for 30 min at 37°C before extracting twice with phenol:chloroform and twice with chloroform alone. The DNA was precipitated with 0.1 volume of 3M sodium acetate (pH7) and 2.5 volumes ethanol at –20°C overnight. The DNA was pelleted and washed as previously outlined, then resuspended in 12µl SMQ.

2.2.5 Ligation into λgt11

A third of the 'adapted' cDNA was used in each of three ligation reactions, set up on ice, together with a 'negative control' using the blunt-end cDNA. 1µg arms were used per reaction in a total volume of 10µl, then 2.5 units T4 DNA ligase added. The

ligation step should be carried out at 16°C, but as there was no way to maintain this temperature the ligation time was increased from the 30min recommended by the protocol. The reaction was kept in an ice bucket as close to 16°C as possible for 1h, then placed in a constant temperature room of 4°C overnight. By the following day, the temperature in the ice bucket was 10°C.

2.2.6 Packaging of the bacteriophage

The whole of each cDNA reaction was then packaged. The Gigapack II kit (Stratagene, USA) and the Amersham λ DNA *in vitro* packaging module were used for the leaf and storage root cDNA libraries respectively. The two parts of the packaging extract supplied were quickly added to the cDNA, gently mixed and incubated at 22°C for 2h. They were then diluted to a final volume of 500µl in SM buffer (Salt-magnesium buffer; 100mM NaCl, 8mM $MgSO_4$, 50mM Tris-HCl pH7.5, 0.01% gelatin [v/v from a 2% stock]) and 10µl chloroform added to prevent bacterial growth. The libraries were stored at 4°C.

2.3 cDNA library screening

2.3.1 Plating cells preparation

The following *E. coli* bacterial host strains were used to prepare plating cells: ER 1647 for the λMOS*Elox* cDNA library and Y1090r$^-$ for the λgt11 cDNA library. The appropriate bacterial strain was streaked out on LB medium–agar plates (0.8–1% [w/v] agar), containing 50µgml^{-1} tetracycline or 50µgml^{-1} ampicillin respectively. These were grown overnight at 37°C for the ER 1647 and 42°C for the Y1090r$^-$. A single colony was used to inoculate a 10ml culture of LB medium supplemented with 50µgml^{-1} of the appropriate antibiotic, 10mM $MgSO_4$ and 0.4% maltose (LBM). This was grown overnight at 37°C on a rotary shaker, then 1ml of the culture added to 50ml LBM. The cells were incubated with vigorous shaking at 37°C until they reached an OD_{600} of 0.5

(approximately 2–3h). They were centrifuged at 717*g* for 5min, then resuspended in 4–10ml of ice-cold 10mM $MgSO_4$. Cells were viable for one week if stored at 4°C, but freshly made cells were usually used.

2.3.2 Titration

20µl of the phage library were mixed with 180µl of SM. Since 100µl of this mix will be plated out this constitutes the 10^2 dilution i.e. that one plaque obtained on a plate of this dilution is equivalent to 10^2 plaque forming units (pfu) per ml of the original phage stock. Serial dilutions of the library were then made in SM buffer up to the 10^7 dilution.

100µl of the appropriate library dilutions were incubated with 100µl fresh plating cells at 37°C. After 15min, 4ml of 0.8% (w/v) top-agar containing 50µgml^{-1} of antibiotic (tetracycline for λMOS*Elox* or ampicillin for λgt11), 0.4% (w/v) maltose, 1mM isopropyl β-D-thiogalactopyranoside (IPTG) and 0.02% (w/v) 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) were added, and the mixture poured quickly onto an LB-amp plate and swirled to cover evenly. Once set, the plates were incubated overnight at 37°C (λMOS*Elox*) or 42°C (λgt11).

The presence of IPTG and X-gal in the agar allows for visual screening of recombinants. Active β-galactosidase expressed by non-recombinants appears blue, while the fusion protein results in white plaques, since foreign DNA is cloned in the *Eco*RI site located within the *lac* Z gene. This method was also used to prepare plates with well-separated plaques for isolating single bacteriophage, or confluent plates using eluted single bacteriophage for DNA preparation by the plate lysate method.

2.3.3 Antibody screening

2.3.3.1 Antibody characterization

2.3.3.1.1 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

Minigels were made for the Protean II kit (BioRad, USA). Each minigel comprised a 10% resolving gel (10% 29:1 [w/v] acrylamide:bisacrylamide mix; 375mM Tris-HCl pH8.8, 0.1% SDS [v/v from a 10% stock], 0.1% ammonium persulphate [v/v from a 10% stock], 0.01% TEMED [v/v]) and a 5% stacking gel (5% 29:1 acrylamide:bisacrylamide mix; 125mM Tris-HCl, pH 6.8, 0.1% SDS [v/v from a 10% stock], 0.1% ammonium persulphate [v/v from a 10% stock], 0.01% TEMED [v/v]). The resolving gels were cast first, in the gel plates, allowed to set and then the stacking gels poured on top and the well-forming comb inserted. For each gel, the acrylamide, buffer and SDS components were mixed in a beaker then the ammonium persulphate and TEMED added to catalyse polymerization. The mixture was rapidly swirled and the gel cast immediately.

Protein from the cultivar MCol 22 had previously been prepared as crude water extract. A small sample of this was taken from –20°C and dissolved in 200µl SDS-PAGE loading buffer (2% SDS [v/v from a 10% stock], 62.5mM Tris-HCl pH6.8, 5% β-mercaptoethanol, 0.0025% [w/v] bromophenol blue, 2% [v/v] glycerol). The sample was boiled for 3min, centrifuged at 9638*g* for 5min and 15µl aliquots loaded in each well of the two minigels. SDS low molecular weight markers (Sigma, UK) were also loaded. The gels were run in 1× Laemmli buffer (25mM Tris, 192mM glycine, 0.5% SDS [v/v from a 10% stock]) until the dye fronts reached the bottom.

2.3.3.1.2 Coomassie staining of SDS-PAGE gels

Gels were stained for 10min with 1% (w/v) Coomassie brilliant blue stain, then destained in 45% methanol:9% acetic acid for 3.5h, all at room temperature.

2.3.3.1.3 Ponceau S staining of SDS-PAGE gels

Membranes to be used for antibody staining which required the positions of molecular weight markers to be indicated were stained at room temperature for 20min with 0.2% (w/v) Ponceau S made up in water, then washed with distilled water.

2.3.3.1.4 Electro-transfer of proteins to nitrocellulose membrane

Following SDS-PAGE, gels were rinsed briefly in transfer buffer (330mM Tris-HCl, 190mM glycine, pH8.4. The nitrocellulose membrane (Hybond-C, Amersham, UK) was cut to the size of the gel, using a pencil mark for orientation, and soaked for 15–30min at room temperature. The gel was then placed within the transfer cassette (BioRad, USA) on a piece of filter paper soaked in transfer buffer. The membrane was placed on top and covered with more filter paper, taking care to ensure no air bubbles were present between any of the layers as they impede even transfer. Transfer was carried out for 1h in transfer buffer at approximately $3Vcm^{-1}$, using the Bio-Ice cooling block to maintain a low temperature.

2.3.3.1.5. Membrane blocking

To prevent non-specific binding of probes or antibodies to membranes, the latter were routinely blocked using a 2% (w/v) solution of either BSA (Sigma, UK) or 'Marvel' milk protein (Premier Beverages, UK) in Tris-buffered saline (TBS; 10mM Tris-HCl pH7.4, 150mM NaCl). The membranes were incubated in solution for 1h at room temperature (or overnight at 4°C) with gentle shaking.

2.3.3.1.6 Primary antibody binding

Membranes with proteins to be detected by antibodies, resulting either from electro-transfer following SDS-PAGE, or direct plate lifts of cDNA libraries, were blocked as described, then incubated shaken for 1h at room temperature with one of the two available primary antibodies; the Brazilian 22kDa protein antibody at the recommended working concentration of 1:2000, and the Bath antibody at a dilution of 1:12.5, the only dilution where a result was obtained.

This was followed by four 5min washes in TBST; TBS plus 0.05% Tween 20 (Sigma, UK) to remove excess primary antibody, before transferring to the goat-anti-rabbit Ig horseradish peroxidase-linked whole antibody (Sigma, UK) diluted 1:1000 in TBS. As previously, incubation was for 1h on an orbital shaker, then excess secondary antibody removed by four further 5min washes in TBST.

2.3.3.1.7 Secondary antibody detection

This was carried out using the SuperScreen immunoscreening system (Amersham, UK), which utilizes diaminobenzidine (DAB) as the chromogenic substrate. DAB was stored at –20°C in 0.4ml aliquots, each sufficient for screening five 9cm circular filters when used diluted in 50ml TBS for library screening (0.2ml was used for the western blots of two membranes). 7.5µl 30% hydrogen peroxide (as supplied) were then added to the diluted DAB and approximately 10ml of the mixture poured immediately over each membrane, turned protein-bound side uppermost. It is important not to shake the filters during signal development, or a diffuse result is obtained. Positive signals develop within 5min and when the optimum signal to noise ratio had been obtained, the reaction was stopped by washing thoroughly with distilled water. As DAB is light-sensitive and unstable at room temperature, developed membranes were stored in the dark and photographed before they faded.

2.3.3.2 Plate lifts of cDNA expression libraries

Circular filters of 9cm radius were cut from Hybond-C nitrocellulose membrane (Amersham, UK) and soaked in a solution of 10mM IPTG before being dried on paper towels. Once dry, the IPTG impregnated filters were placed carefully onto plates with phage plaques and marked asymmetrically. The plates were then inverted and incubated at 37°C for a further 3h. Expression of the *lac* Z gene within the phage is induced by IPTG and in recombinants this results in production of the fusion protein which binds to the nitrocellulose filter. The filters were then removed and washed 3× 5min in TBS to remove any remains of top agar. Fresh IPTG filters were then overlaid onto the plates and incubated at 37°C overnight to create a duplicate set.

2.3.3.3 Antibody screening

Membrane blocking, and incubation with the primary and secondary antibodies were carried out as described above for the antibody characterization.

2.3.3.4 Secondary antibody detection

2.3.3.4.1 DAB

This was again carried out as described above.

2.3.3.4.2 Electro-Chemiluminescence (ECL)

The ECL detection system is around ten times more sensitive than the DAB system. The protocol is the same up to detection of the secondary antibody, with the exception of slightly longer washing times; a single 15min wash followed by two washes of 5min each after blocking non-specific sites and the binding of the primary antibody, and one 15min wash followed by four 5min washes after the binding of the secondary antibody.

An equal volume of reagents 1 and 2 of the ECL kit were mixed together, 0.125ml of final developer mix being sufficient for 1cm^2 of membrane. Excess buffer was drained from the membranes and they were placed on a piece of Saran Wrap (Dow, USA), protein side upwards. The detection reagent was added to the membranes, held by surface tension and incubated for exactly 1 min without agitation. The reagent was then drained off, the membranes blotted and wrapped in Saran Wrap and any air bubbles smoothed out. The membranes were placed, protein side up, in an autoradiography cassette. In the dark, a piece of Hyperfilm-ECL (Amersham, UK) was placed carefully on top and exposed for 15s. This was then removed and replaced by a fresh piece of film. The first film was developed immediately to estimate the correct exposure for the second piece of film (usually around 2–10 min).

2.3.3.5 PCR detection of recombinant bacteriophage

The polymerase chain reaction was used to analyse the percentage of true recombinants within the storage root cDNA libraries and to estimate the size of the clones. PCR was originally tried directly from the phage suspension, since the Amersham protocol advises that this is possible.

Table 2.2 Forward and reverse primer sequences used for PCR to identify recombinants from (i) λ MOS*Elox* and (ii) λgt11 cDNA libraries

Primer name	Sequence (5'–3')
(i) T7 gene 10 primer (forward)	TGA GGT TGT AGA AGT TCC G
T7 terminator primer (reverse)	GCT AGT TAT TGC TCA GCG G
(ii) λgt11 forward primer	GGT GGC GAC GAC TCC TGG AGC CCG
λgt11 reverse primer	TTG ACA CCA GAC CAA CTG GTA ATG

As this was unsuccessful, amplification was repeated after using the polyethylene glycol (PEG) precipitation method of Fraser *et al.*, 1994. Some amplification was obtained, but it was by using the second phage DNA miniprep method below (*2.3.3.5.1*) that amplification was routinely achieved. Table 2.2 lists the various primers used.

Single plaques were cored from agar plates and eluted in 500µl SM buffer at 4°C overnight. 20% (w/v) PEG was added to a concentration of 4%, with 2M NaCl added to a final concentration of 0.5M. The DNA was precipitated on ice for 1h, then centrifuged at 9638*g* for 20min. The supernatant was entirely removed before resuspending the pellet in 100µl sterile water (Fraser *et al.*, 1994).

2.3.3.5.1 Miniprep of phage DNA from agar plaques

Traces of PEG may remain after using the above method, which inhibit PCR, so a second method was used. A single plaque was picked using a sterile pasteur pipette and used to inoculate 5ml LB medium supplemented with 10mM $MgSO_4$. This was grown overnight in an orbital shaker at 42°C, since the host bacterial strain Y1090r$^-$ is temperature sensitive. After the addition of 50µl chloroform, the culture was incubated for a further 15min, then centrifuged at 13461*g* for 5min. The supernatant was removed and stored at 4°C to provide a high titre stock of approximately 10^{11} bacteriophage ml^{-1}. A 600µl portion of this was removed, 6µl 1mgml^{-1} DNase 1 and 1µl 10mgml^{-1} RNase A added to remove bacterial nucleic acids and incubated at 37°C for 1h. Precipitation in 20% (w/v) PEG 8000:2M NaCl in SM buffer for 1h on ice was followed by centrifugation at 7965*g* for 10min. The phage were then resuspended in 600µl resuspension buffer (100mM Tris pH8.6, 300mM NaCl, 50mM EDTA). 10% (w/v) SDS and proteinase K were added to final concentrations of 0.1% and 20µg ml^{-1} respectively, and incubated at 56°C for a minimum of 4h. A further addition of 10µl 10% SDS was made and incubated at 75°C for 15min. After extraction with an equal

volume of 50:50 phenol:chloroform (29:1 chloroform:iso-amyl alcohol [IAA]) saturated with TE buffer, the DNA was precipitated with 0.5M LiCl and 600µl isopropanol for 30min at –20°C. After centrifugation at 13461*g* for 25min as above, the DNA was washed in 0.5ml 70% ethanol, re-centrifuged at 13461*g* for 5min and resuspended in 30µl TE buffer.

2.3.3.5.2 PCR of λMOSElox cDNA library clones

Single white plaques were cored from LB-agar ampicillin-IPTG-X-gal plates and eluted at 4°C in 25µl SM buffer. Amplification was carried out using 5µl eluted phage which was first heated at 99°C for 5min to kill the plating cells and inactivate nucleases. The PCR mix contained 1× reaction buffer (100mM Tris-HCl pH9, 500mM KCl), 1mM $MgCl_2$, 200µM dNTPs, 5pmol each T7 gene 10 and T7 terminator primers and 1 unit *Taq* DNA polymerase (Pharmacia, USA) in a total reaction volume of 50µl. After overlaying each reaction with mineral oil, the samples were amplified using the following cycle on a Hybaid Omnigene thermal cycler (Hybaid, UK): 5min denaturation at 94°C; 35 cycles of 1min at 94°C, 1min at 52°C, 2min at 72°C; 6min at 72°C for a final extension. Samples were cooled to 4°C and stored until analysed.

2.3.3.5.3 PCR of λgt11 cDNA library clones

To isolate clones A24–30, amplification was carried out on 10µl phage DNA template in a total volume of 25µl containing 1× reaction buffer (100mM Tris-HCl pH9, 500mM KCl), 1.5mM $MgCl_2$, 200µM dNTPs, 25pmol each forward and reverse primers and 1 unit *Taq* DNA polymerase (Pharmacia, USA). Reactions were overlaid with mineral oil as before, using the same thermal cycler. PCR conditions were 90s denaturation at 94°C, followed by 30 cycles of 94°C for 1min, 52°C for 1min, 72°C for 1min and a final extension at 72°C for 5min. 15µl of the amplified mix was analysed on a 1% (w/v) agarose gel as described below (see ***2.4.1.1***).

on a 1% (w/v) agarose gel as described below (see ***2.4.1.1***).

Clones E28, E30, Z5, Z9 and Z11 were amplified using the following protocol. A total reaction volume of 25µl containing 2µl phage DNA, 1× NH_4 reaction buffer (16mM $(NH_4)_2SO_4$, 67mM Tris-HCl pH 8.8, 0.01% Tween-20; Bioline, UK), 3mM $MgCl_2$, 200µM dNTPs, 25pmol each forward and reverse primers and 1 unit *Taq* DNA polymerase (Bioline, UK) was overlaid with mineral oil and amplified using a PTC-100 Programmable Thermal Controller (MJ Research Inc, USA). PCR conditions were 90s denaturation at 94°C, followed by 30 cycles of 94°C for 30s, 50°C for 30s, 72°C for 2 min and a final extension at 72°C for 5 min. All the reaction mix was loaded on a 1% (w/v) TAE preparative gel (see ***2.4.1.2***), to allow complete recovery of the DNA for successful amplifications. The DNA was then excised and purified as described in **2.4.2**.

2.4 cDNA clone analysis

2.4.1 Agarose gel electrophoresis

2.4.1.1 TBE gel electrophoresis

For a 1% (w/v) gel, 1g of agarose was weighed in an Ehrlenmeyer (conical) flask and 100ml of either 1× or 0.5× Tris-Borate-EDTA (TBE) buffer added. Usually 1× TBE was used, but for small minigels and when gels were only run for a short time, 0.5× TBE provided sufficient buffering capacity. A mark was put on the outside of the flask to show the original level of the meniscus and the suspension heated gently in a microwave oven until all the agarose had dissolved. Regular swirling ensured no agarose crystals remained and that the mixture did not superheat. The volume of the mixture was adjusted to the original level with more water, ethidium bromide added to a

final concentration of 0.5μgml^{-1} and the gel mix swirled gently to ensure even distribution. It was then poured into the casting tray, a comb with the required amount of teeth placed in to the mix to form the wells, and the gel allowed to set at room temperature.

Loading buffer (final concentration 1×: 0.25% [w/v] bromophenol blue, 0.25% [w/v] xylene cyanol, 15% [w/v] ficoll) was added to all samples including any markers, which were made up to the same volume as the samples to ensure uniform migration. This provides a visual marker to indicate the extent of migration and a weighting substance to allow loading of the samples into the wells. The gel was placed in the gel tank, completely submerged in 1× TBE buffer and an electric current applied, generally 5Vcm^{-1}, and the samples electrophoresed until the bromophenol blue dye front had almost reached the bottom of the gel. The migration of the samples was viewed by illumination with UV light (320nm) as described previously (***2.2.1.3***).

2.4.1.2 TAE gel electrophoresis

This was used preferentially when the DNA being electrophoresed was to be cut out and purified, since DNA binds less strongly to gels buffered with 1× TAE (Tris-Acetate-EDTA). Gels were cast, loaded and run in the same manner as for gels buffered with TBE, except that the running buffer was 1× TAE.

2.4.2 DNA gel band purification

2.4.2.1 Electroelution

Following agarose gel electrophoresis, the band containing the DNA to be purified was cut out of the gel using a new scalpel blade. It was placed inside a short length of No.2 wide dialysis tubing, previously autoclaved in 10mM EDTA and rinsed with TE, and 200μl TE buffer added. The dialysis tubing was sealed ensuring no air

bubbles and the gel slice manoeuvred to the edge. This was then placed inside a minigel tank containing 1× TBE and run at $5Vcm^{-1}$ for around 15 min.

When viewed under UV light at 320nm, the DNA stained with ethidium bromide was seen to have moved out of the gel slice and in to the surrounding TE buffer. This was removed to a sterile eppendorf tube, the gel slice discarded and 100µl TE used to wash the tubing. This was added to the tube containing the DNA and extracted with an equal volume of phenol:chloroform. The DNA was precipitated at –70°C for 30 min with 0.1 volume of sodium acetate (pH7) and 2 volumes of 100% ethanol, then centrifuged for 30 min at 13461*g* and resuspended in a suitable volume of SMQ. The concentration was determined by electrophoresing a small sample in an agarose gel as usual.

2.4.2.2 Geneclean III DNA purification

The required DNA band was excised from the agarose gel. This was then weighed and the volume of the gel slice determined, assuming 1g to be equivalent to 1ml. Purification of the DNA was carried out using the Geneclean III kit (Bio 101, USA). Three volumes of NaI were added before incubating at 55°C for 5 min to melt the gel slice. For fragments smaller than 500bp, 0.1 volume of TBE modifier was added to lower the pH of the NaI and hence facilitate precipitation of the smaller fragments.

After complete solubilization, the EZ-Glassmilk was resuspended and added to bind the DNA. For solutions containing less than 5µg DNA, 5µl of the suspension was used, which was increased by 1µl for each further 0.5µg DNA. This was mixed and incubated at room temperature for 5 min, mixing every 1–2 min to ensure the matrix remained in suspension. The EZ-Glassmilk-DNA was pelleted by centrifugation at 13461*g* for 5s at room temperature. The supernatant was removed and the pellet washed three times with 10–50 volumes (200–700µl) of NEW Wash. Each time the

pellet was resuspended by pipetting and then centrifuged as above, before discarding the supernatant. After the final wash, the tube was centrifuged for a few seconds more and the remaining NaI solution removed.

It is recommended that the DNA be eluted with an equal volume of Elution Solution to that of the EZ-Glassmilk added. Up to three elutions may be performed, obtaining approximately 80%, 10–20% and 1% of the total DNA for the first, second and third respectively. However, since 5µl was a very small volume, it was decided to elute the DNA of two successive 7.5µl volumes followed by a final 5µl volume.

2.4.2.3 Sephaglas DNA purification

Bands were excised from the gel after electrophoresis, transferred to an Eppendorf tube and weighed. For quantities of DNA up to 250mg, 250µl Gel Solubiliser from the Sephaglas kit (Pharmacia, USA) was used, while for larger quantities the initial 250µl was increased by $1\mu l mg^{-1}$. This was vortexed vigorously and incubated at 60°C for 5–10 min until the agarose was completely dissolved. The Sephaglas was vortexed to give a uniform suspension and 5µl added for each 250mg gel slice, with a further 5µl for each estimated 1µg of DNA. This was incubated at room temperature for 5 min, vortexing every minute. After incubation, the Sephaglas-DNA was pelleted by centrifugation at 13461*g* for 30s, and the supernatant removed. 40µl wash buffer was added (or 8× the Sephaglas volume), centrifuged again for 30s at 13461*g* and the supernatant removed. This was repeated for a total of three washes. Finally, the DNA was eluted by adding two lots of 20µl elution buffer (or 4× the Sephaglas volume), and centrifuging as before. Care was taken to ensure none of the Sephaglas resin was transferred with the DNA.

2.4.3 Sub-cloning

2.4.3.1 pTAg / pUAg

The inserts within recombinant bacteriophage were amplified by PCR, electrophoresed on 1% (w/v) TAE gels and the DNA extracted from the gel using one of the band preparation methods described above.

The LigATor kit (R&D systems, Abingdon, UK) was chosen, as the 3.8kb plasmid used, pTAg, is designed for cloning PCR products. PCR frequently produces a product with terminal adenosine residues (Clark, 1988) and this vector contains thymine residue 'overhangs' at either side of the cloning site, thereby facilitating ligation. Numerous reactions were pooled prior to band purification to minimize cloning of products containing errors and where sequence reliability was required, multiple clones were sequenced.

Each ligation reaction was set up using 20–50ng DNA. This was adjusted depending on the size of the insert to maintain the recommended ratio of 3:1–1:1 vector:insert. A self-ligated control (no insert DNA) and a reaction using the control 50bp insert supplied were also set up. 1× ligase buffer (20mM Tris-HCl pH 7.6, 5mM $MgCl_2$), 5mM DTT, 0.5mM dATP and 50ng pTAg vector were combined in a 1.5ml tube, to which was added the PCR fragment and sterile water to adjust the volume to 9.5µl. These were mixed, then the tube centrifuged briefly to collect all components at the bottom. 0.5µl T4 DNA ligase was added and mixed gently using a pipette tip. The reaction was incubated at 16°C overnight.

Competent cells supplied were thawed on ice, mixed gently and 20µl cells carefully pipetted in to a pre-cooled tube on ice, one aliquot per transformation. To these, 1µl of the ligation reaction was added and tapped gently to mix. The tube was incubated on ice for 30min, then the cells heat shocked by incubating at 42°C for exactly 40s. The reaction was then incubated on ice for 2min after which 80µl SOC

medium (2% [w/v] bacto-tryptone, 0.5% [w/v] bacto-yeast extract, 10mM NaCl, 2.5mM KCl, 10mM $MgCl_2$, 10mM $MgSO_4$, 20mM glucose) was added. This was incubated for 1h at 37°C in a rotary shaker at 200–250rpm.

LB-agar plates containing 50μgml^{-1} ampicillin and 15μgml^{-1} tetracycline were spread with 20μl 5% (w/v) X-gal and 100μl 100mM IPTG. After drying for 30 min at 37°C between 25μl and 50μl of each transformation reaction (20μl for the controls) were spread on each plate. After absorption they were incubated at 37°C for 16h.

2.4.3.2 Transformation using TSS competent cells

A single bacterial colony was used to inoculate 10ml LB and grown shaken at 37°C overnight. In turn, 1ml of this was used to inoculate 100ml LB and grown under the same conditions until the OD_{600} was approximately 0.4. The culture was then centrifuged at 4500*g* at 4°C, the pellet resuspended in 10ml ice-cold TSS (LB supplemented with 10% [w/v] PEG 6000, 5% [w/v] dimethylsulphoxide [DMSO] and 25mM $MgCl_2$) and filter sterilized through a 5μm filter. Aliquots of the cells were snap-frozen in liquid nitrogen and stored at –70°C.

For transformation, the cells were thawed on ice. 100μl cells were added to the DNA in an ice-cold eppendorf tube (maximum 10μl DNA), then incubated on ice for 5–60 min. For ampicillin resistant plasmids, the transformed cells were then centrifuged briefly, resuspended in 50μl LB and plated out on to an LB plate containing 50μgml^{-1} ampicillin. For bacteria resistant to other antibiotics, prior to plating out 900μl LB should be added, the cells shaken at 37°C for 45 min, centrifuged, resuspended and then plated onto the appropriate plate.

2.4.4 Single colony gel analysis

Cells from the colonies to be tested were resuspended in 150µl of single colony gel buffer (2% [w/v] ficoll, 1% SDS [v/v from a 10% stock], 0.012% [w/v] bromophenol blue, 0.012% [w/v] orange G) and incubated for 15min at room temperature. The suspension was centrifuged at 13461*g* for 15min, then 40–100µl of the supernatant loaded in the wells of a non-submerged 1% (w/v) agarose gel buffered in 1× TAE. A current of $5Vcm^{-1}$ was applied until the bromophenol dye front was approximately 1cm from the wells. The gel was then submerged in 1× TAE running buffer, the wells thoroughly washed out and electrophoresis continued until the dye front was two-thirds the way down the gel. The gel was viewed under UV illumination as described previously.

2.4.5 Plasmid DNA miniprep methods

2.4.5.1 Alkali lysis miniprep

A single bacterial colony was used to inoculate 2ml LB and incubated at 37°C overnight in an orbital shaker. 1.5ml of this culture was centrifuged at 13461*g* for 3min, the supernatant removed and the pelleted cells resuspended in 0.5ml STE buffer. They were then centrifuged as before, the supernatant removed and the cells resuspended in 100µl ice-cold Solution I (50mM glucose, 25mM Tris-HCl pH8, 10mM EDTA). To this was added 200µl freshly prepared Solution II (0.2M NaOH, 1% SDS [v/v from a 10% stock]) and the preparation mixed gently by inverting several times. 150µl of ice-cold Solution III (3M potassium acetate pH7, 11.5% [v/v] glacial acetic acid) was added, mixed and incubated on ice 3–5 min. After centrifugation at 13461*g* for 5min the supernatant was transferred to a fresh tube and extracted with an equal volume of phenol:chloroform. The supernatant was again removed, two volumes of ethanol added to precipitate the DNA and vortexed. Centrifugation for 5min at 13461*g*

pelleted the DNA, which was then washed with 1ml 70% ethanol and re-centrifuged. Finally the DNA was resuspended in 50µl SMQ containing 20mgml^{-1} DNase-free RNase A.

2.4.5.2 Wizard miniprep

For the Wizard miniprep method (Promega, USA), again a single bacterial colony containing the plasmid and DNA of interest was used to inoculate 5ml of LB overnight at 37°C. Depending on the growth of the culture, 1–5ml of this was centrifuged for 5min at 7965*g* and the supernatant removed. This protocol also employs alkaline lysis of the bacteria but has modifications.

The cells were resuspended in 250µl Cell Resuspension Solution (supplied with the kit) and vortexed. 250µl Lysis Solution (again supplied with the kit) were added, mixed by inversion then incubated until the suspension cleared. 10µl alkaline protease solution were added and the tube inverted several times to mix. This was incubated for 5min at room temperature, after which 350µl Neutralization Solution from the kit were added, mixed and the lysate was centrifuged for 10min at 13461*g* at room temperature.

The supernatant was transferred to a Wizard spin column and centrifuged for 1min at 13461*g*. The flow-through was discarded, 750µl Column Wash added to the column and centrifuged again for 1min at 13461*g*. The flow-through was again discarded, 250µl Column Wash added and centrifuged for 2min at 13461*g*. The column was transferred to a clean Eppendorf tube and the plasmid eluted by centrifuging at 13461*g* for 1min, with 100µl nuclease-free sterile water.

2.4.6 DNA Sequencing

2.4.6.1 Di-deoxy chain termination sequencing

DNA was prepared by alkaline lysis. 15μl were incubated at 37°C for 1h with 1μl of a 1μgml^{-1} solution of RNase A in a total volume of 20μl if RNA was still present. The DNA was further purified by use of a Sephadex G50 column (Pharmacia, USA). The Sephadex was shaken well and 750μl poured in to the column. The packaging stop was removed from the column and the lid placed on, but not too tight. The column was centrifuged at 717*g* for 1min to pack the resin and the eluate removed. The Sephadex equilibrated with 750μl 10mM Tris-HCl pH7.5, vortexed to completely resuspend it, and centrifuged again at 717*g* for 1min. This step was repeated. 5μl of 2M NaOH was added to 20μl of DNA, loaded onto the column and centrifuged at 717*g* for 2min.

Sequencing reactions were carried out using the Sequenase 2 kit (USB, USA). If reactions were being carried out on DNA not immediately after its elution from the column, it was first denatured by heating to 65°C for 5min. For each reaction 7μl of denatured DNA was mixed with 2μl 5× Sequenase buffer, 1μl of the appropriate primer (approximately 5pmol) and made up to a total volume of 10μl with sterile water. This was mixed, pulse centrifuged and incubated at 65°C for 5min. It was then allowed to cool slowly (15–30min) to room temperature to allow the primer to anneal. The mixture was centrifuged briefly to collect all of the sample and chilled on ice.

Aliquots of 2.5μl of each termination mix, containing di-deoxyadenosine, di-deoxycytosine, di-deoxyguanosine and di-deoxythymidine (ddA, ddC, ddG and ddT) respectively were dispensed into microfuge tubes and incubated until required at 37°C. The labelling mix was diluted 1:15, since this allows for longer (though fainter) reactions. The Sequenase enzyme was diluted 1:8 and the following mix prepared: 1μl 0.1M dithiothreitol (DTT), 2μl of the diluted labelling mix, 0.5μl α-^{35}S-dATP (5μCi). This

was added to the ice-cold DNA mix and incubated 2–5 min at room temperature.

3.5µl of labelling reaction were added to each termination tube and incubated for 5 min at 37°C. The reactions were stopped by adding 4µl stop solution.

Samples were denatured for 5 min at 75°C and electrophoresed at 35Vcm^{-1} on a 6% polyacrylamide gel (v/v from a 40% stock) buffered in 0.5× TBE. Short runs were stopped when the bromophenol blue dye had just run off the end of the gel, while longer runs allowed the xylene cyanol dye to migrate until it was 10 cm from the bottom of the gel.

After electrophoresis, the plates were removed from the tank, the back plate removed and the gel soaked for 30 min in solution (5% [v/v] methanol, 5% [v/v] acetic acid). It was then removed, a piece of 3mm Whatman paper cut to size placed on the exposed side of the gel and carefully peeled off. It was covered in Saran Wrap and placed in an autoradiograph cassette with film at –70°C for 3–7d, then developed.

2.4.6.2 Automated sequencing

DNA to be sequenced was one of three types: 1: PCR product; 2: sub-cloned within a plasmid vector; and 3: cloned in a λ vector. Prior to sequencing, the DNA was purified. For PCR products, TAE gel electrophoresis was followed by band purification. Sub-cloned samples in both plasmids and λ vectors were cleaned using Wizard minipreps (Promega, USA). Samples were checked by electrophoresing a small aliquot on an appropriate percentage agarose gel and the concentration determined by comparison with a range of known standards or spectrophotometry.

Dye terminator automated sequencing was carried out using an ABI 377 automated DNA sequencing machine (PE-Applied Biosystems, UK). Initially, 30–90ng of PCR product DNA or 300–500ng plasmid DNA were used, together with 3.2pmol primer made up to a total volume of 12µl with SMQ. This volume was later reduced to 6µl and the components adjusted accordingly. However, sequencing of λ samples

remained in a 12µl volume, with 3.2pmol primer and using around 900ng DNA for each reaction.

2.4.7 Southern analysis

2.4.7.1 Genomic DNA extraction

DNA was prepared from young leaves of the cultivars CMC 40 and MCol 1684 according to the protocol of Dellaporta *et al.* (1983). Approximately 7–8g of tissue were collected, the leaves rinsed in distilled water, and frozen immediately in liquid nitrogen. They were ground to a fine powder and transferred to tubes for centrifugation, approximately 4g per tube. 15ml Dellaporta extraction buffer (100mM Tris-HCl pH8, 50mM EDTA, 500mM NaCl, 0.07% [v/v] β-mercaptoethanol, 1% [v/v] polyvinylpyrrolidone) were added, prewarmed to 40–50°C. 1ml 20% (w/v) SDS was added and mixed well, then incubated shaken in a water bath at 65°C for 10min. The tube was inverted every 2min. 5ml 5M potassium acetate pH7 were added, mixed vigorously then incubated on ice 40–50min. The extraction was then centrifuged at 15000rpm (Sorvall RC5B centrifuge, SS34 rotor) for 20min at 4°C. The supernatant was carefully filtered through a double thickness of sterile gauze into a centrifuge tube containing 10ml isopropanol, precooled at –20°C, mixed gently and incubated at –20°C for a minimum of 2h or overnight to precipitate the DNA. It was centrifuged at 15000rpm for 15min at 4°C as above, the supernatant discarded and the pellet dried at room temperature. The DNA was redissolved in 700µl TE buffer, using a sterile, cut blue pipette tip. Pellets which did not dissolve easily were heated for 5–10min at 65°C. The DNA was transferred to a fresh Eppendorf tube and centrifuged at 9368*g* for 10min at room temperature, then transferred to another fresh tube containing DNase-free RNase to give a final concentration of 10µgml^{-1} and incubated at 37°C for 30min. 75µl 3M sodium acetate (pH7) and 500µl isopropanol, precooled to –20°C, were

added, mixed gently and incubated at –20°C for a minimum of 2h or overnight.

After centrifugation at 13461*g* for 15min at 4°C the supernatant was discarded. 500µl 70% ethanol, precooled at –20°C, were added to wash the DNA, then centrifuged at 13461*g* for 2min at 4°C. This washing step was repeated, then the pellet was vacuum dried for 10–15min. The pellet was dissolved in 200–400µl SMQ and the integrity of the DNA checked on a 0.8% (w/v) agarose gel. The DNA was stored at 4°C.

2.4.7.2 DNA hybridization

Genomic leaf DNA (5–10µg) from the cultivars CMC 40 and MCol 1684 was singly digested with around 40 units of the restriction enzymes *Eco*RI and *Hind*III overnight at 37°C. Digests of gDNA from the bacterial strain XL1Blue were used as controls.

The samples were electrophoresed overnight at 30V in a 0.7% (w/v) agarose gel, buffered in 1× TBE and stained after electrophoresis with a solution of 0.5µgml^{-1} ethidium bromide in 1× TBE. The gel was photographed while under UV illumination (320nm), with a ruler at one size for calculation of fragment sizes. After trimming to size, it was then soaked for 15min in 0.25M HCl, 30min in denaturing solution (1.5M NaCl, 0.5M NaOH) and 30min in neutralizing solution (1.5M NaCl, 0.5M Tris-HCl pH 7.2, 1mM EDTA). Two fibre pads were then soaked in 20× SSC (Standard sodium citrate; 3M NaCl, 88% [w/v] sodium citrate, pH7). Six pieces of 3mm Whatman filter paper were cut to approximately the same size as the gel, with a further two cut to exactly the same size. Seven of these were also soaked in 20× SSC, as was a piece of Hybond-N membrane (Amersham, UK) again cut to exactly the size of the gel. To aid orientation, one corner of the gel and the corresponding corner of the membrane were removed.

The fibre pads were placed in the centre of a tray and six of the filter papers

placed on top. The gel was placed on top of these, with the membrane on top of the gel. After ensuring there were no air bubbles, Saran Wrap (Dow, USA) was placed over the tray, cut along three sides along the edges of the membrane and folded back, keeping the remainder taut. The final wet filter paper, of exact gel size, was placed on the membrane and finally the dry filter paper on top.

Paper towels were then placed on this to a height of around 10cm, then a flat piece of glass or perspex to allow for even weight distribution, and finally the weight (usually a bottle of 500ml solution). This was left overnight at room temperature to allow for transfer of the DNA to the membrane.

The following morning, the weight, all towels and the top two filter papers were removed. The gel with the membrane still attached was carefully turned over and a pencil used to mark the position of the wells on the membrane. The membrane was then briefly washed in 2× SSC, then dried for 15min in a folded piece of filter paper. To fix the DNA the membrane was then baked at 120°C for 20min.

Filters were either probed immediately or frozen at –70°C until required.

2.4.7.3 Probing of membranes

Membranes were briefly rinsed in 2× SSC, then incubated in prehybridization buffer (265mM $Na_2HPO_4.2H_2O$ pH7.2, 7% SDS [v/v from a 10% stock], 1% [w/v] skimmed milk powder) in an orbital shaker at 60°C for at least 1h. One Ready-To-Go labelling reaction (Pharmacia, USA) per probe was used and reconstituted in 20µl SMQ on ice. The probe DNA (10–50ng) was denatured for 10min at 95°C then snap-cooled on ice for 2min. The DNA was briefly centrifuged to ensure no loss of sample and added to the labelling mix. 2.5µl of α-^{32}P-dCTP (around 25µCi) was added, the tube placed inside a lead pot and incubated at 37°C for 15min.

The probe was then denatured for 10min at 95°C, snap-cooled on ice for 2min, 50µl SMQ added and the tube centrifuged briefly to collect all of the sample. The

probe was added to the prehybridization mixture containing the blot and incubated overnight in the shaker at 60°C. The following day, the blot was washed twice with 2× SSC, firstly for 15min, then 30min. The final wash was 15min in sterile distilled water. All washes were at 60°C to ensure stringent binding. The membrane was dried between sheets of 3mm Whatman paper, then covered with Saran Wrap and placed in an autoradiograph cassette with film (Kodak, USA) at –70°C for 1–7d. Films were developed using the X-Omat developing machine (Kodak, USA).

2.5 Genomic clone isolation and analysis

2.5.1 Genomic library screening

2.5.1.1 Membrane production

The cassava genomic library used was the kind gift of Professor Monica Hughes (Newcastle University), made from cultivar MBra 534 as cited in Liddle *et al.*, 1997, and cloned in the vector EMBL3 (Stratagene, USA). The library titre was $2.3 \times 10^8 \text{pfuml}^{-1}$, and the number of colonies required to be screened to have a 99% probability of finding a single copy gene is given by the following equation:

$$\text{Number} = \frac{\ln\ [1 - \text{probability}]}{\ln 1 - \left[\frac{\text{insert size}}{\text{haploid genome size}}\right]}$$

where: the probability of finding a single copy gene = 0.99
average insert size for EMBL3 = 12–15kb
cassava genome size = 772Mb,

hence: $$\text{Number} = \frac{\ln\ [1 - 0.99]}{\ln 1 - \left[\frac{15}{772\,000}\right]}$$

$$= \textbf{237 010 plaques}$$

From these calculations it was decided to plate out 250000 plaques on each of four 22 x 22cm square plates, with four positive clones expected.

A total of 4µl of the library was diluted in 800µl SM buffer. The *E. coli* strain LE392 was used as a host for EMBL3 and 2ml plating cells were used together with 200µl of the library dilution per plate. After incubating at 37°C for 15min, this was mixed with 50ml molten top-agarose and poured onto a pre-warmed plate. After allowing them to set, the plates were incubated at 37°C overnight.

The plates were placed in the cold room (4°C) for 1h, then overlaid with duplicate membranes (Hybond N^+; Amersham, UK), marked at the edges with an asymmetric pattern using a sterile needle. The first membrane was overlaid for 30s and the second for 90s. Each was rinsed for 30–60s in denaturing solution, 5min in neutralizing solution, then rinsed briefly in 2× SSC. After drying on blotting paper the DNA was fixed on the membranes by baking at 120°C for 15min.

Probing of the membranes was carried out in the same manner as for Southern analysis (***2.4.7.3***).

2.5.2 Large scale preparation of phage DNA

2.5.2.1 Plate lysate of phage DNA

Confluent plates were made as described above from single recombinant bacteriophage. These were then covered with 3ml λ diluent (10mM Tris-HCl pH7.5, 10mM $MgSO_4$), shaken and the phage eluted at room temperature for 1–2h. The diluent was then pipetted carefully and centrifuged at 4000rpm for 10min at 4°C (Sorvall RC-5B refrigerated ultracentrifuge, SS34 rotor) to remove any agarose. 1µg each of RNase A and DNase I were added and incubated for 15min at 37°C.

An equal volume of 20% (w/v) PEG 8000:2M NaCl in λ diluent was added, vortexed and incubated on ice for a minimum of 1h. This was then centrifuged at

10000rpm for 10min at 4°C (Sorvall ultracentrifuge as above) and the supernatant removed. The pellet was resuspended in 0.5ml TE buffer and vortexed, after which 5µl 10% (w/v) SDS was added and incubated for 5min at 68°C. 10µl 5M NaCl was added, then the DNA extracted with an equal volume of 50:50 buffer-saturated phenol:chloroform (29:1 chloroform:IAA) followed by extraction with an equal volume of 29:1 chloroform:IAA alone.

The aqueous phase was removed to a fresh tube, an equal volume of isopropanol added and mixed, then incubated at –70°C for 15min. This was then centrifuged at 11470*g* for 15min at 4°C, the DNA pellet washed in 70% ethanol, recentrifuged as above for 5min and allowed to air dry at room temperature. Finally the DNA was resuspended in TE buffer or sterile water and stored at 4°C.

2.5.2.2 Liquid lysate of phage DNA

Positive plaques were used to make high titre stocks, using the plate lysate method and eluting in a volume of 5ml per plate. Two 10ml cultures of NZYCM were inoculated with 10µl LE392 plating cells and incubated in a shaking incubator at 37°C overnight. 400µl of each of these were then used to inoculate each of two flasks containing 100ml NZYCM and incubated shaken at 120rpm, at 37°C until the culture reached an OD_{600} = 0.6 (approximately 3h). To each was added 500µl plate lysate, which was again incubated shaken overnight at 37°C for lysis to occur. The cultures were then placed on ice overnight in the cold room before being pooled.

RNase A and DNase 1 were added to a final concentration of 1µgml^{-1} and left at room temperature for 2h. NaCl was then added to a final concentration of 1M and incubated on ice for 1h. The culture was then centrifuged at 11000rpm for 15min at 4°C (Sorvall RCB5; SS34 rotor) and the supernatant retained. 10% (w/v) PEG 8000 was added to this and slowly dissolved using a magnetic stirrer at 4°C. This was then precipitated on ice overnight.

After centrifuging at 11000rpm for 15min at 4°C as above, the supernatant was discarded and any remaining fluid removed. The pellet was dissolved in 4ml SM buffer, an equal volume of 29:1 chloroform:IAA added and vortexed for 30s. This was then centrifuged at 3000rpm for 15min at 4°C and the aqueous phase removed to a new tube. EDTA was added from a 0.5M stock (pH8.0) to a final concentration of 20mM, proteinase K to 50$\mu g ml^{-1}$ and SDS to 0.5%, both final concentrations, then mixed by inversion and incubated at 56°C for 1h.

The preparation was cooled to room temperature and an equal volume of phenol equilibrated with Tris-HCl (pH8.0) added. Again it was mixed by inversion until an emulsion was formed, the centrifuged for 10min at 3000*g*. The aqueous phase was twice extracted with an equal volume of 50:50 phenol:chloroform (29:1 chloroform:IAA) and once with an equal volume of 29:1 chloroform:IAA. To the final aqueous phase was added 0.1 volume of 3M sodium acetate (pH7.0) and 0.7 volume isopropanol, and the DNA precipitated for a minimum of 15min at –20°C (usually overnight).

After centrifuging at 13461*g* for 10min, the DNA pellet was rinsed in 70% ethanol and spun briefly. The supernatant was removed, the pellet air dried for 5min and then resuspended in 200µl SMQ at 4°C overnight. Storage was also at 4°C.

CHAPTER 3: NUTRIENT STUDIES ON CASSAVA ROOT FORMATION

3.1 INTRODUCTION

3.1.1 Potential for manipulation of protein content in the storage root

Despite the low amount of protein within the cassava storage root, experiments have shown that some proteins are more abundant than others, in particular a protein from the parenchyma of 25kDa (Clowes *et al.*, 1994). Since the tuber storage proteins studied by other research groups (such as patatin in potato and sporamin in sweet potato) are within the 20–40kDa molecular weight range, this looked a promising candidate for investigation. Furthermore, the biological roles of patatin and sporamin, and the probable derivation of dioscorin from an enzyme, may well indicate that storage proteins are selected by the plant on a basis of their presence in the appropriate location, regardless of the original function. Hence it may be feasible to stably up-regulate, introduce or recruit a protein in a similar manner into the cassava storage root. In addition, alteration of the amounts of nitrogen and sulphur available to the young plants forming storage roots appeared to increase the amount of the 25kDa protein (Clowes *et al.*, 1994).

3.1.2 The biological roles of nitrogen and sulphur

Nitrogen and sulphur play important roles in plant growth and development and are regarded as essential mineral nutrients. While they are needed for a number of cellular components, their fundamental role is in protein synthesis, nitrogen being found in all amino acids and sulphur present in methionine, cysteine and cystine. Nitrogen is also a component of all nucleotides. Thus, any plant lacking either of these elements

shows retarded growth and chlorosis, or yellowing of the leaves (Taiz and Zeiger, 1991). Other characteristics associated with nitrogen-deficiency are for the plants to possess thin, woody stems and it has also been observed that carbohydrates not used in nitrogen metabolism can be re-routed into the production of anthocyanins, resulting in purplish stems, leaves and petioles. Cassava has a remarkable ability to grow on poor soils (Cock, 1985) and it is perhaps not surprising that these latter characteristics are to be seen in the crop.

Sulphur has also been shown to be important in the quality of bread made from wheat (*Triticum aestivum* L.), increasing the gel protein content but decreasing the elasticity (Zhao *et al.*, 1997) and it is possible it may have a similar role in cassava flour.

3.1.3 Effect of altering nitrogen and sulphur supply to the storage root

Experiments investigating the effect of supplying increased nitrogen and/or sulphur to plants on their root formation had suggested that nitrogen deficiency limited root size but not number, and that sulphur deficiency resulted in smaller roots with lower dry matter content (Clowes *et al.*, 1994). Furthermore, increasing the available nitrogen resulted in an up-regulation of the 25kDa parenchyma protein of interest. It was therefore decided to verify these studies with a larger sample size. Of the two Colombian cultivars MCol 1684 and CMC 40 used previously, slightly more nitrogen had been found in the roots of the former (Clowes *et al.*, 1994), so it was chosen for use here. Four different nutrient solutions; high nitrogen–low sulphur (HN–S), high nitrogen–high sulphur (HN+S), low nitrogen–high sulphur (LN+S) and low nitrogen–low sulphur (LN–S) were used to assess the response of the plants.

3.2 RESULTS

3.2.1 Visual characteristics

The appearance of the plants grown under the different regimes varied. Low sulphur, regardless of the amount of nitrogen supplied, resulted in paler leaves with a more yellow appearance. In contrast, the leaves and pre-lignified stems of plants with high sulphur were deep green and their leaves were larger. These differences can be seen in Figure 3.1 for the HN–S and HN+S plants.

Figure 3.1 Plants grown with high nitrogen supply, (A) without and (B) with additional sulphur.

3.2.2 Extent of root formation

Roots produced by the plants in each of the groups also varied. More roots were produced in total by the plants given low sulphur; 23 for LN–S and 20 for HN–S compared with 14 for LN+S and 12 for HN+S. The number of roots produced by individual plants, together with their individual and cumulative wet weights are shown in Table 3.1, and were used for the subsequent statistical analysis.

Table 3.1 Numbers and wet weights of roots produced by the six individual plants for each nutrient regime

	LN–S		LN+S		HN–S		HN+S	
	No. of roots	Wt (g)*	No. of roots	Wt (g)*	No. of roots	Wt (g)*	No. of roots	Wt (g)*
Plants 1–6								
Total:	23	2345.77	14	3178.78	20	3226.10	12	913.68
Average:		101.99		227.06		161.31		76.14
Plant:								
1	2	27.44	1	**566.20**	1	**407.43**	0	–
		217.48						
		244.92						
2	2	113.65	2	166.45	2	193.23	1	**35.24**
		278.81		328.94		268.85		
		392.46		**495.39**		**462.08**		
3	3	163.69	2	331.32	2	290.73	2	102.46
		155.89		517.17		175.90		69.53
		144.15		**848.49**		**466.63**		**171.99**
		463.73						
4	4	34.73	3	3.12	4	90.84	2	21.70
		147.26		31.79		129.92		356.31
		180.02		91.91		124.37		**378.01**
		206.66		**126.82**		235.68		
		568.67				**580.81**		
5	6	2.01	3	57.81	5	14.96	3	65.72
		2.94		94.70		40.80		51.25
		15.72		141.87		50.35		40.13
		29.52		**294.38**		142.76		**157.10**
		41.17				290.63		
		149.84				**539.50**		
		241.20						
6	6	16.93	3	125.13	6	40.35	4	10.53
		34.34		161.99		59.22		10.62
		51.46		560.38		61.58		19.43
		63.58		**847.50**		173.10		130.76
		67.58				203.40		**171.34**
		200.90				232.00		
		434.79				**769.65**		

*Total weight of roots in g for each individual plant is highlighted in bold

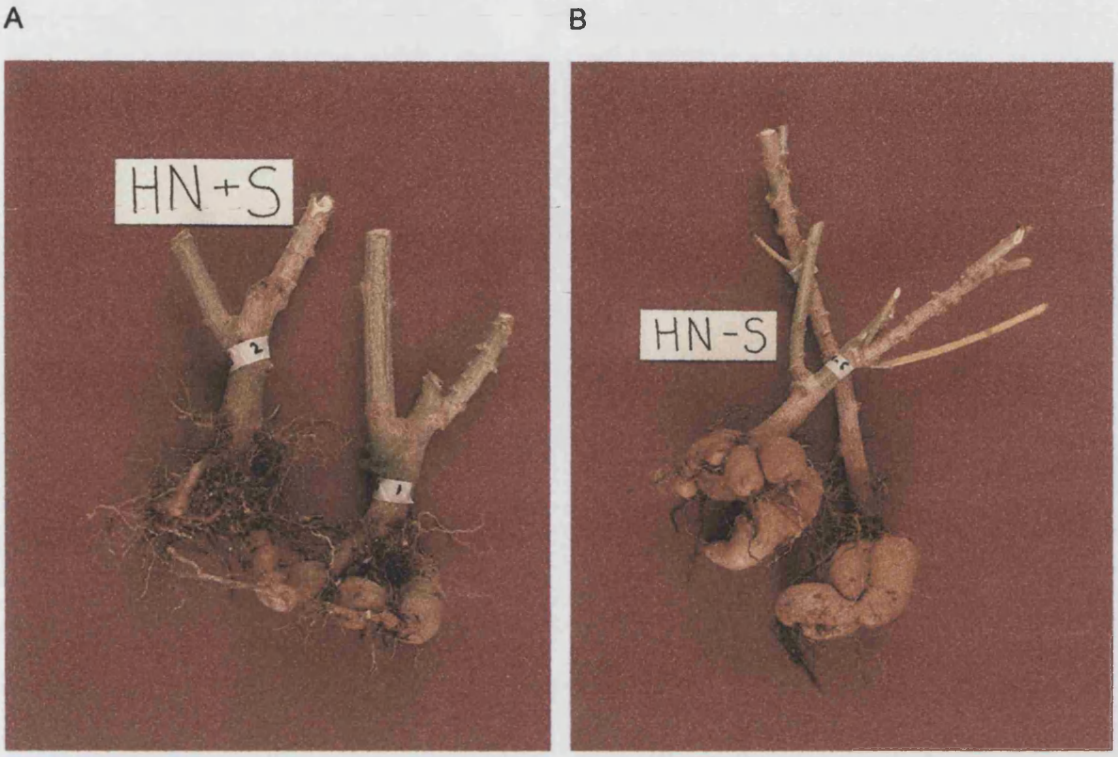
With both nutrients at low levels (LN–S), the number of storage roots produced per plant varied from two to six storage roots. The two plants with six roots each had one that was noticeably larger than the rest. A similar range in root number was observed for the HN–S group, but overall there was an increase in the weight of roots produced per plant, resulting in a total increase of 780g over the LN–S total.

When sulphur was supplied at higher levels, less storage roots were observed per plant, with a maximum of three produced for the LN+S group and four for the HN+S group. Increasing the sulphur increased t he overall yield when nitrogen was low, but when both nutrients were high, yield was markedly decreased. The difference in storage root production with altered sulphur levels when nitrogen was high can be seen in Figure 3.2. The total weight of storage roots produced by the low nitrogen groups increased with high sulphur from 2.3kg (LN–S) to 3.2 kg (LN+S), whereas with the high nitrogen weights decreased from 3.2kg to 0.9kg (all figures to the nearest 0.1kg).

3.2.3 Statistical analysis

Figure 3.3 shows the means of the total weight of roots formed in g per plant, together with standard error bars. There is a marked decrease in the amount of root formation in the treatment containing high levels of both nutrients. The data was analysed by using a 2-way ANOVA, a table for which can be seen in Appendix 3. From this it was found that neither the effect of nitrogen ($F_{1,20} = 2.36$; $P > 0.10$) nor sulphur ($F_{1,20} = 2.69$; $P > 0.10$) alone was significant, but that the interaction was significant ($F_{1,20} = 12.17$; $P < 0.05$).

Figure 3.2 Roots produced on two plants for each of the high nitrogen groups (A) with and (B) without additional sulphur.



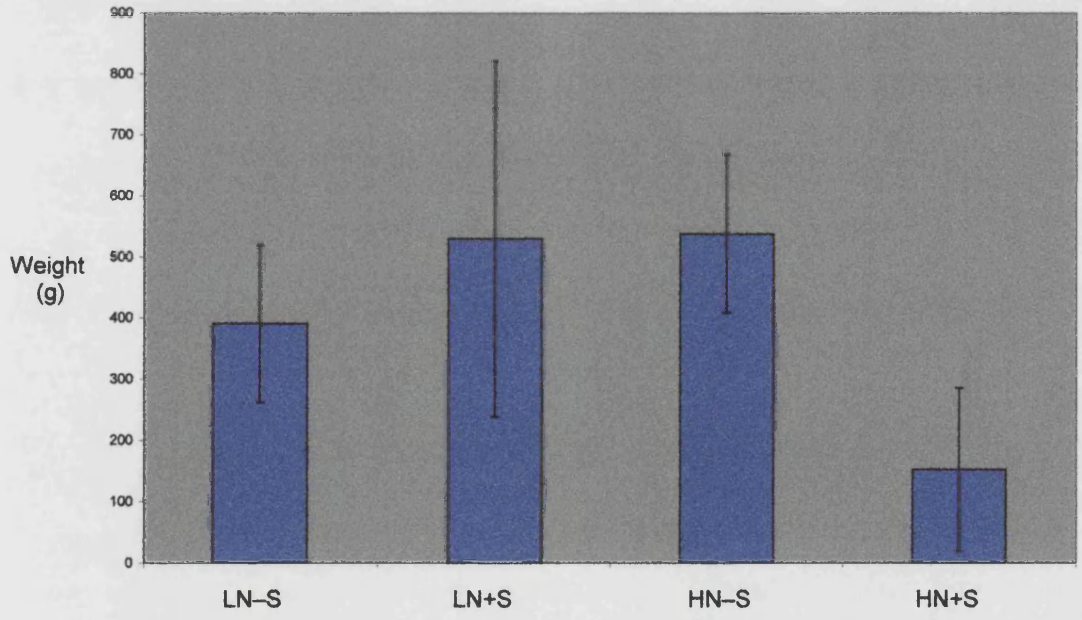


Figure 3.3 Bar chart showing the means of the total weights for roots produced by plants of each of the four nutrient treatments. Error bars are shown, equal to one standard deviation.

3.3 DISCUSSION

3.3.1 Reasons for using fertilizer

Growing cassava on rich soils is generally deemed to be a waste of resources, since it thrives well in adverse conditions, and there is evidence that as wealth increases the choice of cassava as carbohydrate source decreases (Cock, 1985). Thus the question may well be asked "Why bother with studies on the response of the roots to increased nutrient supply?". Cassava is still food for approximately 500 million people around the tropical world (Cock, 1985) and, although associated with some of the world's poorest people in difficult growing conditions, this is not exclusively so. Particularly in Asia, cassava is becoming a crop grown for profit, and information regarding increased yield and possible manipulation of storage root numbers on a plant may be of interest to those who can afford the investment. Research establishments, for example IITA and CIAT, have extensive crop breeding programmes to help local farmers with new or improved planting material. Equally important is the fact that if there is no apparent benefit to be had from applying fertilizer to the crop, then precious resources should not be wasted.

Most studies have concentrated on supplying extra nutrients in the form of nutrient-rich solutions or adding inorganic fertilizers to the soil, but alternatives have been considered. This approach is becoming increasingly popular and may also represent a cheaper alternative in some areas. For example, the use of 'ipil-ipil' (*Leucaena leucocephala*) green manure, supplemented with inorganic potassium and phosphorus has given promising results of increased yield in both cassava and taro (Escalada and Ratilla, 1998).

3.3.2 Adverse effects on root formation of increased nitrogen and sulphur

Altering the levels of nitrogen and sulphur available in the nutrient feed for 24 cassava plants of cultivar MCol 22 had a marked effect on root production when both nutrients were supplied in large amounts. Although neither nutrient was statistically significant when increased by itself, the combination dramatically reduced the yield. The only plant not to produce any tuberous roots was plant 1 of the HN+S treatment block and of the other five plants in this group, four gave storage root weights below 70g. This is probably due to a partitioning effect where aerial growth is favoured over root formation, since over-abundance of nutrients is unlikely to favour the unlimited production of carbohydrate sinks which have no reproductive function and are present solely for times of stress.

Slightly fewer roots were produced by those plants given a higher level of nitrogen, with 20 roots produced by the HN–S plants compared with 23 for LN–S, and 12 roots for HN+S compared with 14 for LN+S. This agrees with earlier observations suggested that lower nitrogen limited the size of storage roots but not the number (Clowes *et al.*, 1994). An increase in mean root weight per plant was observed between LN–S and LN+S (390.96g compared with 529.80g) and between LN–S and HN–S (390.96g compared with 537.68g). However, as can be seen from the summary of the ANOVA in Appendix 3, neither increase was statistically significant ($P > 0.10$ in both cases). However, when both nutrients were increased i.e. HN+S compared with LN–S, a statistically significant decrease in mean root weight per plant was observed (152.28g compared with 390.96g, $P < 0.05$).

Overall, there appears to be some advantage to increasing the level of either nitrogen or sulphur from the low nitrogen environments, since this gave an increase from 2.3kg to 3.2kg in both cases. Experiments at IITA in Nigeria, where cassava was grown in the field, indicated better yields with increased nitrogen fertilizer (Olasantan *et al.*, 1997). Other field trials in Australia, where cassava was grown on post-mined soil

with nutrients replaced in bands, indicate there may well be an optimum level of nitrogen fertilizer above which a negative effect is observed (Fulton *et al.*, 1996), since decreases in both dry matter and harvest index (the ratio of fresh roots to total biomass) were recorded at levels of 200 and 400kgha^{-1} nitrogen fertilizer, compared to those obtained at 0–100kgha^{-1}.

Results from experiments studying glucosinolate concentrations in oilseed rape (*Brassica napus* L.) have indicated strong nitrogen–sulphur interactions (Zhao *et al.*, 1997) and the implication is that application of these fertilizers should be balanced for both elements.

CHAPTER 4: ISOLATION OF CLONES FROM A CASSAVA cDNA EXPRESSION LIBRARY

4.1 INTRODUCTION

Since the aim of this project was to identify gene promoters which would allow manipulation of the root protein quantity and quality, those controlling root-specific or at least root-expressed genes were likely to be of most interest. The 25kDa protein isolated from the parenchyma, and introduced in Chapter 3, had already been used to inoculate a rabbit and produce antiserum, since its size and response to nitrogen made it appear an ideal candidate. Isolating the gene for this protein was the next stage in obtaining a useful promoter.

4.1.1 Use of antibodies in screening cDNA libraries

Once an antibody has been raised to a particular protein of interest, it is then possible to obtain the corresponding DNA sequence, at either the genomic or the expressed level. It was decided to use the antibody raised to the cassava protein described in the previous chapter to obtain the expressed DNA sequence.

The strategy used to obtain an expressed DNA sequence uses the polyadenylated RNA population of the tissue concerned. This is purified from the total RNA and is then used to create a set of complementary DNA (cDNA) sequences to each of the mRNAs present. Thus there should be at least one copy of the cDNA of interest within this population, in this case that corresponding to the 25kDa parenchyma protein discussed earlier. The enzyme used is retroviral reverse transcriptase, which creates a DNA strand complementary to the RNA template, and the initial primer used

can be either a d(T) oligomer (a short stretch of deoxythymidine residues, usually up to 25) which anneals to the poly (A) tail or a mix of random hexamers, priming from sites complementary to them and which may be located anywhere within the mRNA. For subsequent screening by antibodies, the random primer approach is recommended, since there is less likelihood of only obtaining 3' sequence and thus more chance of the epitope to which the antibody was raised being reverse-transcribed. Reviews of these procedures can be found in Alberts *et al.* (1994) and the technical guide to the cDNA synthesis kit (Amersham, 1994a).

Once the first strand cDNA has been made, the mRNA is within the RNA:DNA hybrid is degraded by RNase H activity (processive exonucleolytic removal of rNTPs). DNA Polymerase I then synthesizes the second DNA strand. Each cDNA must then be inserted individually into a vector, the choice of which depends on the size of the cDNA of interest or the average size within the population, and is then referred to as a clone. The cloned population of cDNAs is referred to as a cDNA library.

4.1.2 λMOS*Elox* cDNA library

A previous research student in the laboratory had constructed a cDNA library in the expression vector λMOS*Elox* (Amersham, UK), which is based on the bacteriophage lambda (λ). This accepts cDNA sequences up to 8kb in length and allows for antibody screening by the production of a fusion peptide of the insert with the first 260 amino acids of the phage T7 major capsid protein (Amersham, 1993). Prior to insertion, the ends of the cDNA and the vector must be compatible, achieved in this case by the ligation of specific adaptors to the ends of the cDNA. The adaptors contained restriction sites for the enzyme *Eco*RI, which has a unique site within the vector. To preferentially select for cDNA clones up to 2kb in length, and to remove any remaining unligated adaptors, a size fractionation step was performed and the cDNAs ligated into the vector. To obtain viable phage particles, a packaging reaction was

carried out and the phage was used to infect an appropriate bacterial strain, allowing rapid amplification of the DNA to give sufficient quantities to work with (Amersham, 1994b,c).

A sample of a second antibody, the gift of Dr F Campos, Fortaleza, Brazil, was also available for use, raised to another small cassava storage root parenchyma protein of 22kDa (Souza *et al.*, 1998). The original aim was to use one or other of these antibodies to isolate the corresponding cDNA sequence, by immunoscreening the cDNA library.

4.2 RESULTS

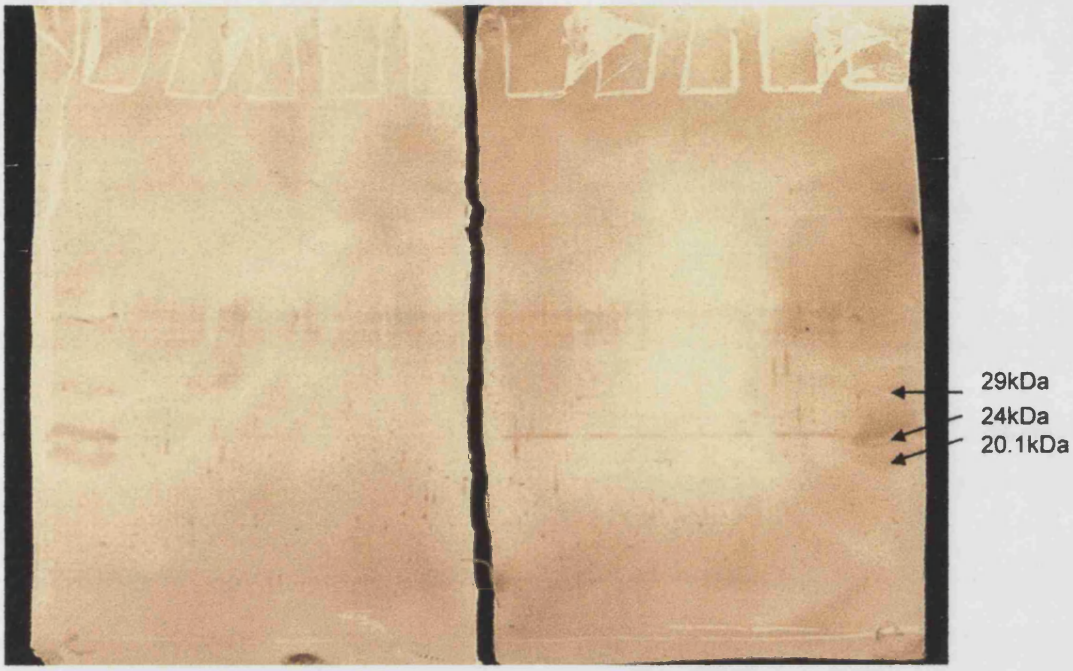
4.2.1 Characterization of antibodies

As both antibodies had been raised to small parenchyma proteins either would be suitable for identifying storage root gene promoters. It was therefore important to characterize them by western blotting.

Figure 4.1 shows the two halves of the nitrocellulose filter, developed using the DAB system. The left-hand part of the filter shows the half probed with the 22kDa Brazilian antibody, while the right-hand side shows the result for the half bound with the 25kDa Bath antibody. Molecular weight markers, seen in pencil in the outermost wells on each side, are indicated.

The Brazilian antibody did not detect any protein, and there was only a small amount of this antibody available, so this was not used further. The Bath antibody was used at a dilution of 1:12.5, which seemed very concentrated compared with the usual expectation of being able to dilute an antibody to around 1:1000, and was cause for concern. However, it was decided to use it to screen the cDNA library, since faint bands were obtained at the expected molecular weight of approximately 25kDa. It was also thought that more target protein might be produced by the expression library.

Figure 4.1 Western blot of cassava total protein using (a) Brazilian antibody to 22kDa protein and (b) Bath antibody to 25kDa protein. Positions of the markers in pencil (Sigma VII Low) correspond to staining of the membrane prior to antibody detection in 0.2% Ponceau S.



4.2.2 Analysis of λMOS*Elox* cDNA library

Six plates of the 10^4 dilution were used to titer the λMOS*Elox* cDNA library. The results obtained were 28, 22, 22, 21, 19 and 17, giving an average of 21.3 plaques per plate. This is equivalent to a library titer of 2×10^5 plaque forming units per ml ($pfu\,ml^{-1}$), which is low. Therefore, before continuing with immunoscreening, it was decided to check out how many of the white plaques were in fact genuine recombinants.

Figure 4.2 shows the results from the PCR of individual white plaques, with clones 12–14 loaded again on the right-hand side of the 100bp ladder to help visualize the sizes of false positives. The size of the amplification product between the primers used, without any insert, is 171bp. With a single *Eco*RI adaptor cloned this increases to 191bp and the product size with two adaptors ligated into the phage is 211bp.

Therefore, any clones for further investigation would need to be visibly larger than this. As can be seen from the photographs, it appeared that all 26 white phage screened using PCR were false recombinants corresponding to one of the three possibilities above. It was therefore decided to construct a new cassava storage root cDNA library.

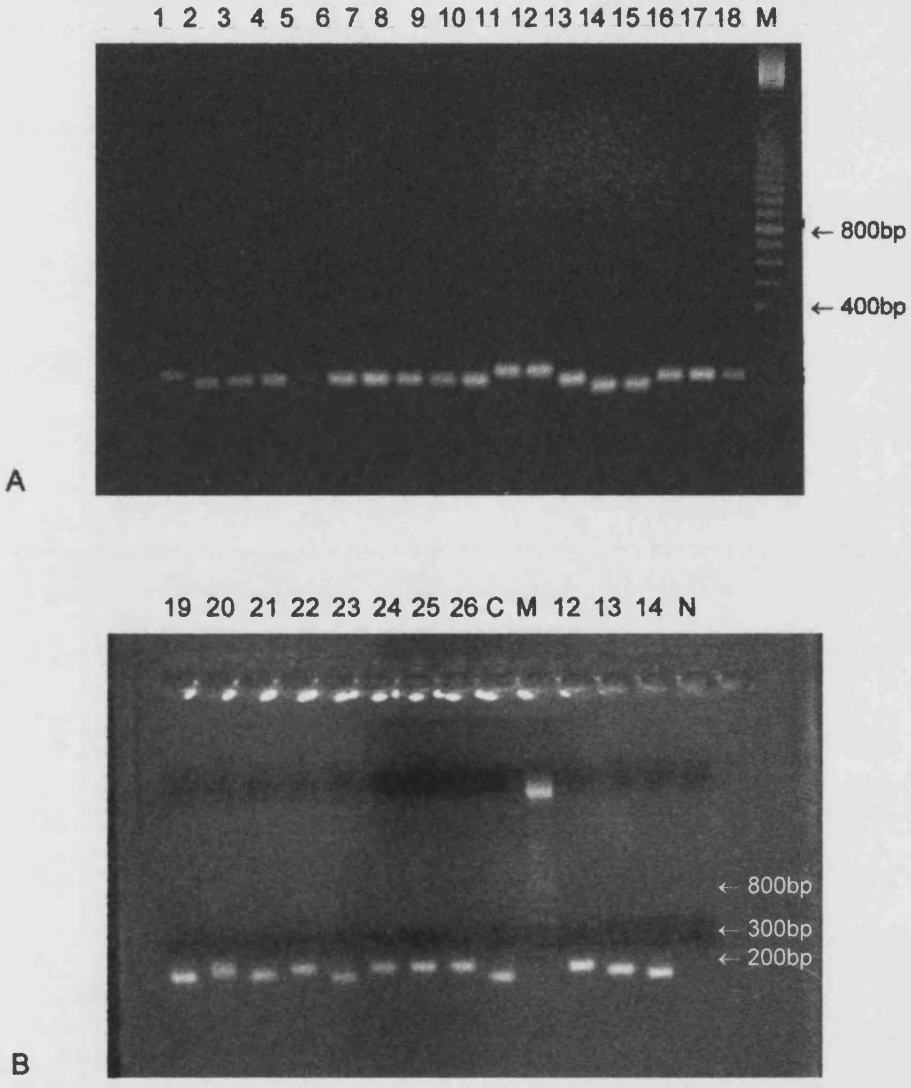


Figure 4.2 Electrophoresis at $5Vcm^{-1}$, in 1% (w/v) agarose gel, of PCR amplifications from the λMOS*Elox* cDNA library, with 100bp ladder DNA molecular weight markers (Pharmacia, USA). A: clones 1–18, M: 100bp ladder; B: clones 19–26, C: λMOS*Elox* control, M: 100bp ladder, clones 12–14, N: negative control.

λMOS*Elox* had not been the easiest vector to work with, requiring the host *E. coli* strain ER1647, which had twice died during storage at the recommended temperature of –20°C, so λgt11 was chosen in preference for the construction of the new library. As with λMOS*Elox*, λgt11 creates a fusion protein using the insert, but this time the bacterial gene utilized is the lac Z gene. An *Eco*RI site is located near the 3' end, into which the foreign cDNA is ligated, producing a β-galactosidase fusion protein in the presence of IPTG and X-gal. This then allows for colour screening of recombinants, since an interruption of the lac Z gene results in the loss of the wild-type blue colour (Amersham, 1994d).

4.2.3 The construction of a new cDNA library in λgt11

Since cassava storage roots contain almost 85% starch, it is extremely difficult to extract RNA from them, and it was decided to firstly construct a cDNA library with leaf tissue since this is easier to manipulate. The subsequent construction of the cDNA library from the more awkward storage root tissue was therefore made simpler by familiarity with the procedure. This in turn gave a better chance of making a successful library, with sufficient recombinants for immunoscreening.

4.2.3.1 mRNA preparation

Good quality RNA is probably the most important criterion for the construction of a representative cDNA library. There should be no degradation and the RNA should be quantified accurately.

Figure 4.3 Total RNA samples from CMC 40 cultivars; A: leaf (lane 1) and B: 3 aliquots of separate storage root samples (lanes 1–3), electrophoresed at $1Vcm^{-1}$ overnight in a 1.2% (w/v) agarose:2.2M formaldehyde gel. The bands corresponding to the 26S and 18S ribosomal subunits are indicated.

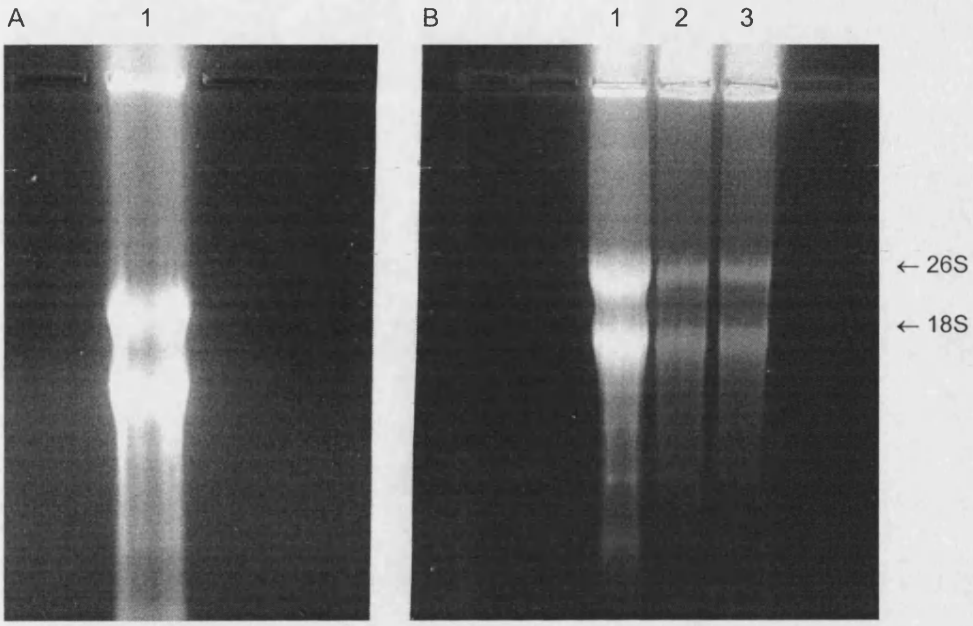


Figure 4.3 shows gels of the respective total RNA preparations for both leaf and storage root. Ribosomal genes occur as repeated units, occurring many times in plants (Alberts *et al.*, 1994). The 26S and 18S ribosomal subunits therefore have many transcripts and appear as discrete bands, and are used as indicators that the RNA is of good quality. The 26S band should appear brighter than the 18S band as it is longer and therefore binds more ethidium bromide and hence fluoresces more brightly. This indicates there has been no significant degradation, while a smear throughout the lane is taken to indicate the presence of differing sized transcripts. Smaller bands visible in lane 1 of the leaf sample gel may correspond to the 5.8S transcript and other repeated transcripts.

Various RNA preparations were pooled to create the final samples for library construction, with 1.2mg total RNA for the storage root library. After purification on Qiagen Tip-100s, the RNA samples were quantified and qualified. This purification was a costly step, reducing the root RNA yield to about 400μg. The successive rounds of

chromatography resulted in a poly (A^+) fraction of approximately 1.5μg. As mRNA is assumed to contribute 1% to the total RNA population, a yield of 0.37% of the original total RNA value is lower than expected but within the limits of experimental error.

4.2.3.2 cDNA library construction

Table 4.1 shows the results from the incorporation of ^{32}P during cDNA synthesis of the storage root library, all calculations being those given in the accompanying booklet to the Amersham cDNA synthesis module.

$$\text{1st strand incorporation} = \frac{\text{cpm B}}{\text{cpm A}} \times 100\% \qquad (1)$$

$$\text{Control incorporation} = \frac{\text{cpm CB}}{\text{cpm CA}} \times 100\% = \frac{3}{236} \times 100\% = \mathbf{1.27\%}$$

$$\text{cDNA incorporation} = \frac{\text{cpm RB}}{\text{cpm RA}} \times 100\% = \frac{5}{298} \times 100\% = \mathbf{1.68\%}$$

Table 4.1 Cerenkov counts measuring ^{32}P incorporation.

Sample	Counts per min (cpm)*
Blank	8
Control unwashed 1st strand filter (CA)	236
cDNA unwashed 1st strand filter (RA)	298
Control unwashed 1st strand filter (CB)	3
cDNA unwashed 1st strand filter (RB)	5
Control unwashed 2nd strand filter (CC)	534
cDNA unwashed 2nd strand filter (RC)	421
Control unwashed 2nd strand filter (CD)	18
cDNA unwashed 2nd strand filter (RD)	5

* All experimental values are the result of the actual value obtained minus that for the blank

2nd strand incorporation = $\frac{\text{cpm D} - (\text{cpm B}/5.2)}{\text{cpm C} - (\text{cpm B}/5.2)} \times 100\%$ (2)

Control incorporation = $\frac{18 - 3/5.2}{534 - 3/5.2} \times 100\%$ = **3.25%**

cDNA incorporation = $\frac{5 - 5/5.2}{421 - 5/5.2} \times 100\%$ = **0.96%**

The value of the second strand reaction is 104µl, of which 20µl is the first strand reaction. This means that the first strand reaction is 1/5.2 of the volume of the completed second strand reaction and therefore cpm B must be divided by 5.2.

Amount of DNA synthesized = (% incorporation x 140)ng

% incorporation = Y% = 3.25% control

= 0.96% storage root

Amount of unlabelled dCTP in 20µl first strand reaction and 104µl second strand reaction mixes = 10nmol

Amount of unlabelled dCTP incorporated = $\frac{Y}{10}$

Total amount of dNTPs = $4 \times \frac{Y}{10}$ (for all 4 dNTPs)

Assuming that the residue molecular weight of 1 mole dNMP = 350ng:

Wt of cDNA synthsized = $350 \times \frac{(4 \times Y)}{10}$ ng

Yield of control cDNA = $350 \times \frac{(4 \times 3.25)}{10}$ = 455ng

Yield of storage root cDNA = $350 \times \frac{(4 \times 0.96)}{10}$ = 134.4ng

Starting from 1μg mRNA, the expected conversion to cDNA is between 20 and 50%, i.e. 0.2–0.5μg cDNA. Greater than 90% of this would then be expected to be transcribed into second strand cDNA, i.e. 0.18–0.45μg cDNA. The value obtained for the control template are slightly higher than this, which is not surprising, and the value for the cDNA from the storage root is in the expected range for the lower first-strand conversion percentage.

4.2.3.3 Titration of the storage root cDNA library

Table 4.2 shows the number of plaques obtained from titrations of the leaf and storage root cDNA libraries (white/blue indicating recombinants/non-recombinants) from duplicate plates in all cases.

Although there are high numbers of non-recombinants in both cases, the storage root library has a greater proportion of recombinants than the leaf library and the overall titer is higher. When the three aliquots of each library (R1, R2 and R3) have been pooled, the titer for the leaf is 1.5×10^{4}pfuml^{-1}, while the storage root is 3×10^{5}pfuml^{-1}.

Table 4.2 Titrations of the leaf and storage root libraries. Values are for the number of plaques obtained on duplicate plates of the serial dilutions and give the numbers of white/blue (i.e. recombinant / non-recombinant) where these could be counted

	10^6	10^5	10^4	10^3	10^2
Leaf					
Control	24 / –	166 / –	1660 / –		
	32 / –	302 / –	1660 / –		
R1	– / –	– / –	1 / –	6 / 12	40 / 51
	– / –	– / –	– / 2		
R2	– / –	– / –	3 / 1	20 / 14	103 / 88
	– / –	– / –	3 / 3		
R3	– / –	– / –	– / 1	4 / –	32 / 23
	– / –	– / –	– / –		
Root					
Control		107 / –	1000s / 14	Confluent	
R1		4 / –	39 / 7	317 / 121	
		2 / 3	40 / 6	367 / 107	
R2		4 / 1	30 / 13	273 / 102	
		6 / 2	33 / 10	316 / 112	
R3		5 / 1	41 / 10	339 /130	
		3 / 1	27 / 14	325 / 93	

4.2.4 Immunoscreening of the storage root cDNA library

Despite extensive screening with both the DAB and ECL systems, no positive plaques were identified with the antibody to the 25kDa protein. Some of the ECL films did show what appeared to be positives, but unfortunately, when compared with the plates, they did not align with plaques. Duplicate filters screened with the DAB detection method also failed to give positive signals in the same area.

Crude tests were done to determine which step of the procedure was causing the problem. Spots of the secondary antibody pipetted directly onto membrane in a recognizable pattern were detected by the ECL system. Primary antibody, incubated as

usual with secondary antibody then detected by ECL, also gave a positive result. However, when total cassava protein was spotted on to pieces of membrane, there was no detection using either signal development method. Figure 4.4 shows film for the ECL system developed from the three membranes: protein spots at the top, primary antibody in the middle and secondary antibody at the bottom.

This result appears to indicate that the primary antibody is not recognizing the protein extract, although it appeared to do so (albeit to a limited extent) in the western blots. It was decided therefore to concentrate on screening the storage root library by PCR, in an attempt to isolate a number of clones from it, some of which may be interesting and possibly root-specific.

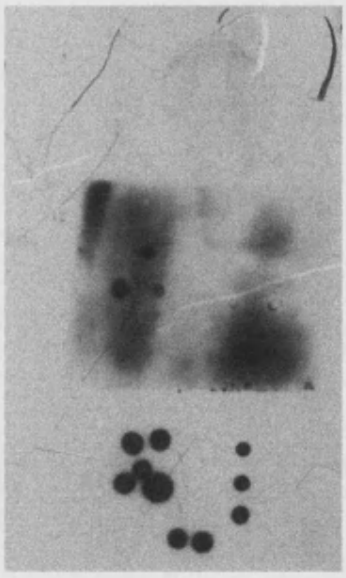


Figure 4.4 ECL film of the test to detect the three different components of the development system, developed for 30s.

4.2.5 PCR screening of the cDNA library

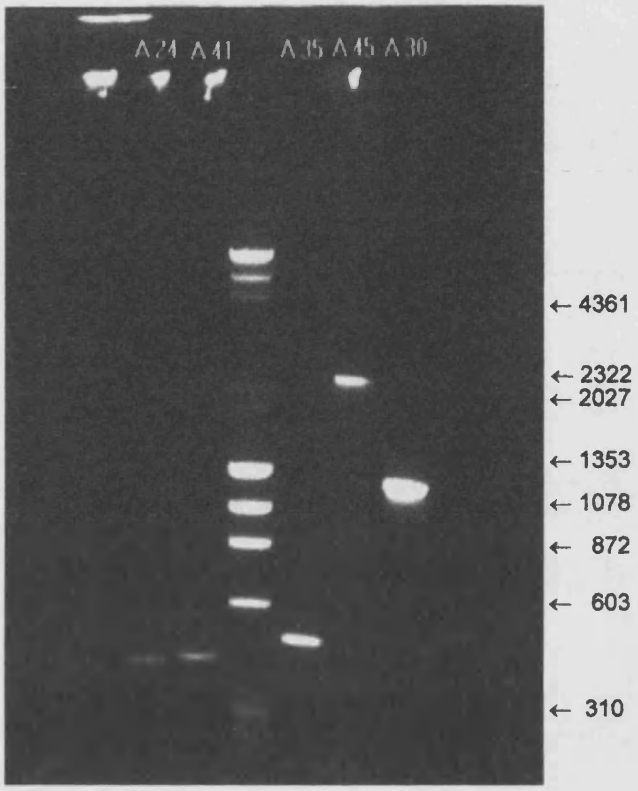
Using the protocol advised by the Amersham booklet, it is possible to amplify products of up to 2kb, which would allow identification of partial and possibly even complete clones for cDNAs of interest. Initial problems of amplification were overcome by using a PEG precipitation step (Fraser *et al.*, 1994). This in turn requires careful removal of all traces of PEG, as it inhibits PCR.

Initially, no amplification could be obtained. However, after optimizing the PCR procedure by adjusting the Mg^{2+} concentration, five recombinant bacteriophage were identified out of 52 white plaques from the library (all of which would have been expected to contain inserts). The schematic diagram in Figure 4.5 shows the relative positions of the λgt11 forward and reverse primers used, the *Eco*RI adaptors used in cloning and the intervening stretches of vector DNA sequence. PCR products from bacteriophage containing cDNA inserts would therefore be of a size greater than 125bp and the five clones obtained are shown in Figure 4.6.

Another five clones were later identified from 48 white plaques. Figure 4.7 shows the PCR products for clones E28 and E30, isolated by a project student, while Figure 4.8 shows three PCR products for clones named as Z5, Z9 and Z11.

Figure 4.5 Position of cDNA inserts relative to the flanking λgt11 sequences. F: forward λgt11 primer and forward *Eco*RI adaptor; R: reverse λgt11 primer and reverse *Eco*RI adaptor.

Figure 4.6 Electrophoresis, in 1% (w/v) agarose gel at $5Vcm^{-1}$, of the first five clones obtained from recombinant λgt11. Molecular weight marker (M) is the λ *Hin*dIII–φX-174 *Hae*III digest (Pharmacia, USA), with fragment sizes in bp indicated down the right hand side.



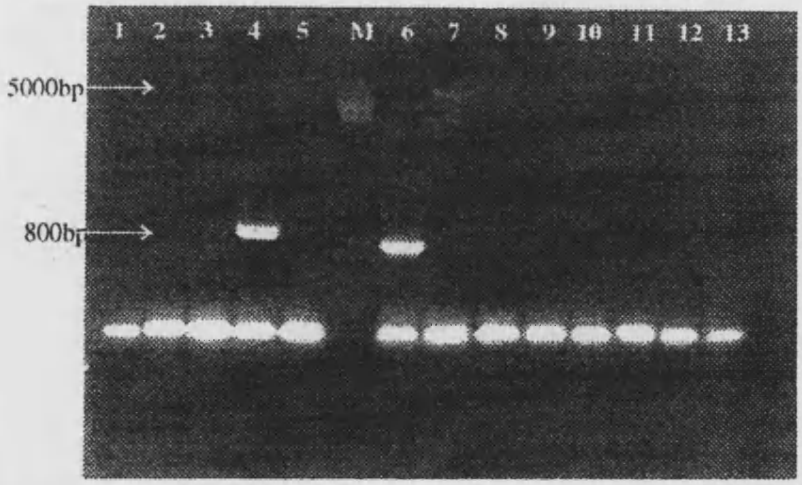
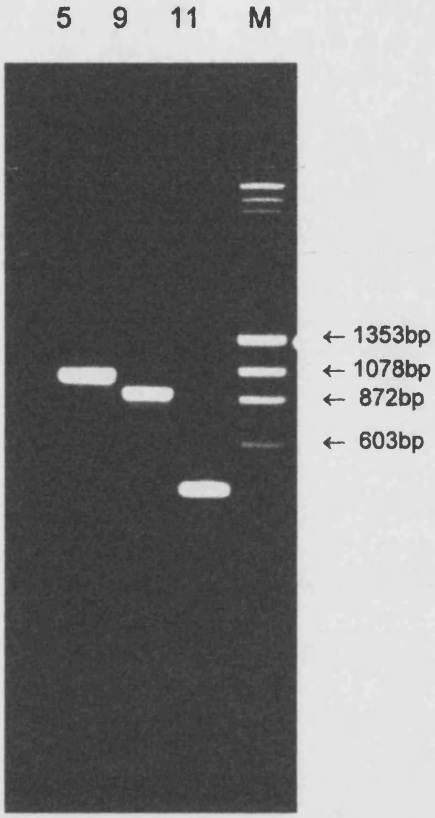


Figure 4.7 Electrophoresis at $5Vcm^{-1}$ in 1% (w/v) agarose of PCR products from clones 1–13. The molecular weight marker is the 100bp ladder (Pharmacia, USA), with the brighter band at 800bp indicated. Lane 4 contains E28, while lane 6 contains E30.

Figure 4.8 Electrophoresis at $5Vcm^{-1}$ in 1% (w/v) agarose, of clones Z5, Z9 and Z11, with molecular weight markers (lane M) the λ *Hin*dIII–ϕX-174 *Hae*III digest (Pharmacia, USA).



4.3 DISCUSSION

Two antibodies were available for immunoscreening and initially it was expected to characterize one or both of the proteins corresponding to these antibodies, using the parenchyma storage root cDNA library in λMOS*Elox*. However, as this library did not appear to have the frequency of inserts required to make antibody screening viable, the alternative was to construct another cDNA library. Despite experiencing many problems with both antibodies, it was decided to proceed using the expression vector λgt11 to allow for screening with the Bath antibody, which hybridized weakly with a 25kDa protein.

After construction of the new library, it was not subsequently possible to detect protein using the Bath antibody. This meant that either the antibody or the library was faulty, or possibly both. It was therefore imperative to determine whether anything had in fact been cloned in the new library or whether to abandon it. At this point strategies for the detection of root-specific clones such as differential display were considered. However, although it was possible to go to Long Ashton Research Station in order to do this, it is not an easy technique to master and with one year already elapsed, the decision was made to at least check the library for inserts first. Because few cDNAs had been isolated from cassava (Hughes and Hughes, 1994; Hughes *et al.*, 1994; Salehuzzamen *et al.*, 1992, 1993) this strategy was likely to yield clones of interest.

The ten clones isolated were a variety of sizes and are all discussed in the following chapters.

CHAPTER 5: ANALYSIS OF SIX CLONES; A35, A45, Z5, Z9, Z11 AND E28

5.1 INTRODUCTION

Ten clones had been identified from the cDNA library, as discussed in the previous chapter. The next step was to identify the sequences, to select the most interesting for further study, and to verify that they were indeed from cassava. One or more clones could then be chosen to characterize fully at the genomic level. This chapter deals with six of the clones; A35, A45, Z5, Z9, Z11 and E28.

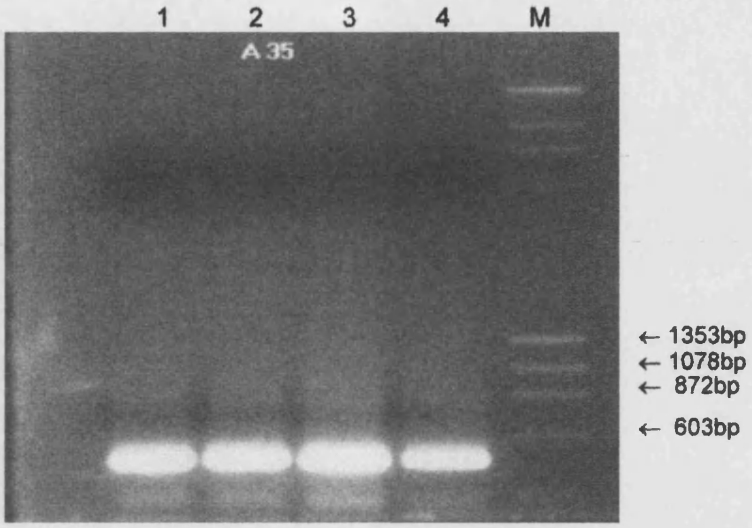
5.2 RESULTS

5.2.1 A35

The PCR product for A35 was approximately 450bp, 125bp of which was the various flanking sequences and primer sites from the λgt11 vector (as shown in Figures 4.5 and 4.6). This remaining insert was 325bp.

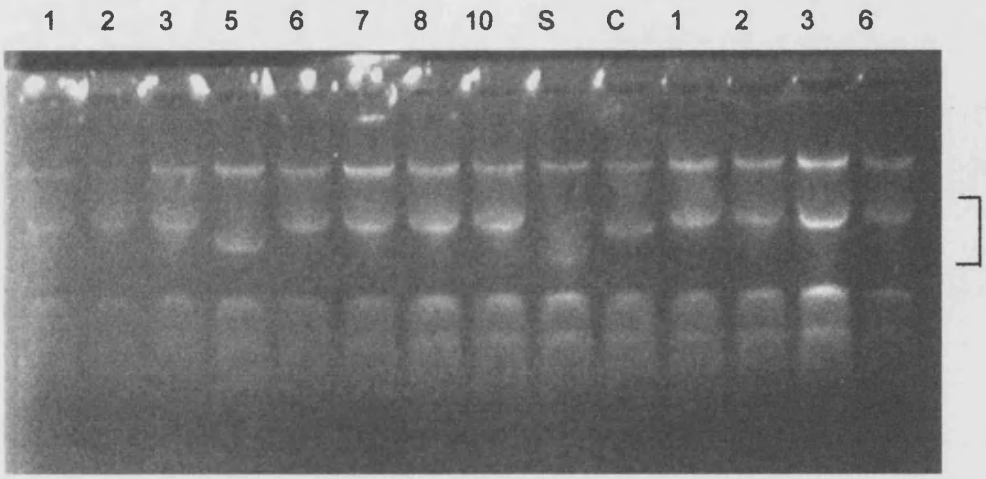
Identification was by the di-deoxy chain termination method of sequencing (Sanger *et al.*, 1977), which required the sub-cloning of A35. This in turn required reamplification of the PCR product (Figure 5.1). The bands were then excised, pooled and the DNA purified from the gel. After checking a small sample on another gel to allow determination of the concentration as $4ng\mu l^{-1}$ (not shown) A35 was sub-cloned into the pTAg vector (R&D Systems, UK).

Figure 5.1 Gel electrophoresis of the A35 PCR products using λgt11 primers (lanes 1–4) in 1% (w/v) agarose, at 5Vcm^{-1}, using 1× TAE as a buffer. Markers (M) were λ DNA *Hin*dIII/ϕX-174 DNA-*Hae*III digest (Pharmacia, USA).



Blue-white selection allowed the positive colonies to be identified. However, since white colonies are not always genuine recombinants, a number were tested by the rapid method of single colony gel electrophoresis (Figure 5.2). Compared with the self-ligated negative control, all white colonies except number 5 were recombinants. DNA was prepared from these positives by the miniprep alkaline lysis method, then used for sequencing. Since PCR products had been used, it was important to obtain reliable sequence from a number of sub-clones of A35, as there is an estimated frequency of mis-match incorporation of bases *in vitro* of 2.6×10^{-5} during PCR (Loeb and Kunkel, 1982).

Figure 5.2 Single colony gel analysis of white colonies from the sub-cloning of A35. S: self-ligated pTAg control; C: pTAg with control 50bp insert. Band variation was observed in the area indicated by the bracket.



Three sub-clones were sequenced initially, in both directions, using the λgt11 primers. More sequence was obtained from the forward reaction, with 230bp, while only 60bp were obtained from the reverse reaction. However, this was sufficient to compare with the GenEMBL (European Molecular Biology Laboratory) and HGMP (Human Genome Mapping Project) databases using the FASTA and BLAST alignment programmes (Pearson and Lipman, 1988; Altschul *et al.*, 1990).

Figure 5.3 shows the alignment for the partial sequence of A35. This clone was most similar to the *E. coli* cDNA sequence for the molybdate binding protein (modA). In some areas with multiple bases of the same nucleotide, it was unclear from the autoradiograph film whether, for example, there were three or four bases. Such examples appear in lower case. However, assuming the number of repeated bases to be the same as for the *E. coli* sequence, 100% identity over 220bp was obtained. This would not be expected between two such diverse organisms as *E. coli* and cassava, which implied that it is more likely to be a sequence derived from the bacterium than a cassava homologue.

Figure 5.3 The partial nucleotide sequence for A35, compared with the *E. coli* molybdate binding protein (Mod A) cDNA. Lower case letters indicate bases that are unclear on the autoradiographs, but possible.

```
A35:        TGTTGGTACgGTACTGGCtTTTGCTCGTTCTCTCGGTGAGTTTGGTGCAACC
            ||||||||| |||||||| ||||||||||||||||||||||||||||||||
Mod A:      TGTTGGTACGGTACTGGCTTTTGCTCGTTCTCTCGGTGAGTTTGGTGCAACC

A35:        ATCACCTTTGTGTCGAACATTCCTGGTGAAACGCGAACCATTCCTTCTGCCA
            ||||||||||||||||||||||||||||||||||||||||||||||||||||
Mod A:      ATCACCTTTGTGTCGAACATTCCTGGTGAAACGCGAACCATTCCTTCTGCCA

A35:        TGTATACCCTGATCCAGAcCCCCGGCGGGGAAAGTGGAGCGGCGAGACTGTG
            |||||||||||||||||| ||||||||||||||||||||||||||||||||
Mod A:      TGTATACCCTGATCCAGACCCCCGGCGGGGAAAGTGGAGCGGCGAGACTGTG

A35:        CATTATTTCTATTGCGCTGGCGATGATCTCCCTGTTGATTTCAGAATGGCTG
            ||||||||||||||||||||||||||||||||||||||||||||||||||||
Mod A:      CATTATTTCTATTGCGCTGGCGATGATCTCCCTGTTGATTTCAGAATGGCTG

A35:        GCCAGAATCAGG
            ||||||||||||
Mod A:      GCCAGAATCAGG
```

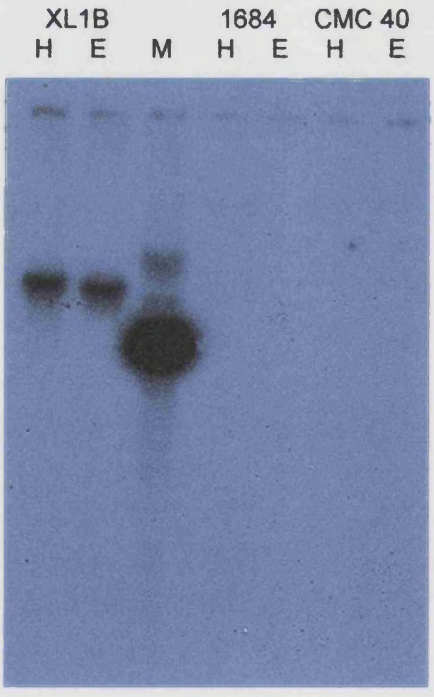


Figure 5.4 Autoradiograph of Southern blot for 10μg CMC 40 and 5μg MCol 1684 total gDNA samples singly digested with *Eco*RI (E) and *Hin*dIII (H), and probed with ^{32}P-labelled PCR product of A35. A total gDNA preparation from *E.coli* XL1B, similarly digested, provided a control. The molecular weight marker (M) was the λ *Hin*dIII digest. Exposure time at –70°C was 7d.

Southern analysis for A35 (Figure 5.4) confirmed that it was not a cassava sequence. Hybridization was obtained with the control bacterial sample but not with either of the two cassava gDNA preparations, despite a 7d exposure time. A35 was therefore not investigated further.

5.2.2 A45

Attempts were made to sub-clone A45 in the same way as A35 but unfortunately this did not work. This may have been because larger fragments are more difficult to clone and A45 was approximately 2.4kb. However, the department had recently bought an automated ABI sequencer (PE Biosystems, UK) and it is possible to sequence directly from PCR products using this automated fluorescent dye method, which allows simultaneous analysis of a population of the amplified fragment. This eliminates the problem encountered with sub-cloning from PCR products, that clones containing amplifications with wrongly incorporated bases will be sequenced by chance.

Therefore A45 was sequenced with both λgt11 primers. The reaction with the reverse primer did not work, but 659bp were obtained from the forward run.

Figure 5.5 shows the result of the BLAST search for A45, aligned with the sequence for tolerance to colicins (Tol) from *E. coli*. There is 79.6% homology over 582bp, although this percentage might have been increased with the resolution of those bases denoted 'N' electronically in A45, by checking the chromatograms from the sequencing runs. Unfortunately, these were not available at this early stage of the running of the machine and only the electronic files of the sequencing runs were received.

Figure 5.5 Partial nucleotide sequence alignment for A45 with the *E. coli* tolerance to colicins cDNA (Tol). Unresolved nucleotides from the automated sequencing files are denoted 'N'.

```
A45:   GAAGNATTG-GNNAGAAACTANANGGNGAACAAGAAANGCNNGAAGNAANTTGANANAAGTN
       |||| | || |  |||||| |   |  |||| || || ||  |||| || |||| | |
Tol:   GAAGAACTGCGTGAGAAACAAGCGGCTGAAC-AGGAACGCCTGAAGCAACTTGAGAAAGAGC

A45:   GTTTAACGGTTCNGGAGCAGAAAAAGCCAGGTTGGAGNAGCNACNAAACAGNCAGGGTTAAA
       | ||| ||| || |||||||||||||| |||| || || |||  | |||||| | | ||||||
Tol:   GGTTAGCGGCTCAGGAGCAGAAAAAG-CAGGCTGAAGAAGCCGCAAAACAGGCCGAGTTAAA

A45:   NNAGTAGNANNTTGAAGA-GAAACA-NGAAAGCGGNGGNNAGATGATAAAGGGTAGCGGGAA
          || || |   |||||| |   || |||||||| ||  ||||| ||||||||||   |||
Tol:   GCAGAAGCAAGCTGAAGAGGCGGCAGCGAAAGCGGCGG-CAGATGCTAAAGCGAAGGCCGAA

A45:   GCAGATGCTANANNTGCGGAAGAAGCAGCGAAGAAGGCGGTTGCAGA-GCAAAGAAAAAANG
       |||||||||| |  |||||||||||||||||||||| |||| |||||| ||||||||||| |
Tol:   GCAGATGCTAAAGCTGCGGAAGAAGCAGCGAAGAAAGCGGCTGCAGACGCAAAGAAAAAA-G

A45:   CAGAAGCGGAAGCCACCAAAGCNGCAGCCGAAGCGCANAAAAAAACCGAGGCAGCNGNTGNG
       ||||||| |||||| ||||||| |||||||||||||| |||||| |||||||||| | || |
Tol:   CAGAAGCAGAAGCCGCCAAAGCCGCAGCCGAAGCGCAGAAAAAAGCCGAGGCAGCCGCTGCG

A45:   GCANTGAAGAAGAAGTTGGTGAAAAAGCCAAAGCAGAAGTTGAGAAGAAAGCGGCTGTTNAA
       ||| |||||||||| || |||||||||||||||||| ||||||||||||||||||| ||
Tol:   GCACTGAAGAAGAAGCTGCTGAAAAAGCCAAAGCAGAAGCTGAGAAGAAAGCGGCTGCTGAA

A45:   AAGGTTGCAGNTGATNANAAAGCGACAGCAGAGANAGNTGCAGCCGACNAAAAAGCNGCAGT
       |||| ||||| |||| | |||||| ||||||||| || |||||||||| ||||||| ||||
Tol:   AAGGCTGCAGCTGATAAGAAAGCGGCAGCAGAGAAAGCTGCAGCCGACAAAAAAGCAGCAG-

A45:   NAAAAGCGGTTATTGAAAAGGCAGCAGCTGATAAGAAAGCNGCGGCAGAAAAAGCCGCTGCA
        |||||||| |  |||||||||||||||||||||||||| |||||||||||||||| |||
Tol:   AAAAAGCGGCTGCTGAAAAGGCAGCAGCTGATAAGAAAGCAGCGGCAGAAAAAGCCGCCGCA

A45:   GACNTNGANGCNGCAGCGGCAAAAGATGCAGNTGAAAAAGCNGNTNCAGCNTAAGCGACNGN
       |||    | || |||||||||||||| ||||| ||||||||| | | ||||  ||||| | |
Tol:   GACAAAAAAGCGGCAGCGGCAAAAGCTGCAGCTGAAAAAGCCGCTGCAGCAAAAGCGGCCGC

A45:   NGCGCCAGATG-TATTTTTTGTGA
        | | |||||| ||||||  ||||
Tol:   AGAGGCAGATGATATTTTCGGTGA
```

Figure 5.6 Autoradiograph of Southern blot for 10μg CMC 40 and 5μg MCol 1684 total gDNA samples singly digested with *Eco*RI (E) and *Hin*dIII (H), and probed with ^{32}P-labelled PCR product of A45. XL1B total gDNA was also digested with these enzymes and the marker (M) was the λ *Hin*dIII digest (Pharmacia, USA). Exposure time at –70°C was 3d.

This Southern analysis for A45 shows a similar result to that for A35, in that the probe bound to the bacterial digests with a single band but showed no detectable hybridization to the cassava samples.

5.2.3 Z5

Since the ABI automated sequencing machine enabled direct sequencing to be undertaken, this was the method chosen for identification of the 'Z' and 'E' clones.

Z5 was approximately 1kb in length, but the sequence obtained was limited. Despite obtaining up to 700bp of unambiguous sequence for other fragments during good runs of the ABI machine, only 125bp in each direction of Z5 were readable. This may be due to the sequence being very A-rich, which results in read-through problems. The sequencing reaction using the forward primer ended with a stretch of 32 nucleotides, 21 of which were adenosine (equivalent to 65.6%). Very small peaks were obtained after this point for around 40bp, after which a characteristic end-point was seen.

The run using the reverse primer did give small peaks on the chromatogram

after the first (approximate) 130bp, but these were unclear and not reliable. Again, there was an A-rich region, consisting of 15 adenosine residues within a 19bp stretch (78.9%).

Figure 5.7 shows a number of alignments for short lengths of the Z5 sequence. The alignments with A and B are more reliable since they used sequence within the early part of Z5. C and D used later sequence, with much lower peaks, and therefore genuine but weak signals may have been missed, by them merging into the 'background' level which is always present. In particular, guanidine residues routinely give much lower peaks compared with the other nucleotides. However, there was only limited identity in under half of the sequence obtained in either direction.

Many problems had been encountered with Southern analysis for the various clones, and no result was obtained for Z5. It was anticipated that investigation of this clone could be continued in the future, but as sequencing was also proving difficult and it was unclear whether Z5 represented a message of interest, it was given low priority.

```
Z5:     GCTTTTGCAGGTCCAATGCTTCTATTCAGTGTGTTTGCTTTTGTTCTTTTTTCCCCAT
        | |||||     |    ||  || |||   ||| ||||||||||||| | |||| || ||
A:      GTTTTTGGTTATTGTATTTTTTTATGTTGTGAGTTTGCTTTTGTTTTGTTTTACCTAT

Z5:     GGAAAAAAGAACAAAAGCAAACACACTGAATAGAAGCATTGGACCTG
        |||   |||||||||||||||| |  |||  | || || | || ||
B:      GGACTGAAGAACAAAAGCAAACAAAAGGAAGGGTAGTATGGCACATG

Z5:     TAGGGAGGGGGGAGGTCGATGTAGGAGTGTGGGGGAGAGAGAGGTCGTAGGG
        |||||||||||| || |  ||||| || | | || |  |  ||  ||| ||
C:      TAGGGAGGGGGGATGTTGTAGTAGGCGTCTTGCGGGGGCACGGGATGTACGG

Z5:     CAGGGGGAGATAGGGAGGGGGGAGGTCGATGTAGGAGTGTGGGGGAGAGAGAGGTCGTAGG
        ||| ||| |   |||| ||||| |   |    ||  | | | |||||||||||| || |||
D:      CAGCGGGGGGGGGGGAAGGGGGGGAAGGGGGGAGATGGGGGAGGGAGAGAGAGGACGGAGG
```

Figure 5.7 Nucleotide alignments for part of the sequence obtained for clone Z5. A: *Herpetomonas samuelpessoai* mRNA sequence; B: *Mus musculus* ERCC2 gene; C: *Gallus gallus* (clones 14, 15) complete retrotransposon ART; D: Pseudorabies virus genome, Rsp40 and pk genes.

5.2.4 Z9

This clone yielded a PCR product of approximately 880bp, which did not sequence well. In fact the chromatogram obtained seemed to indicate a mixed PCR product. Although gel electrophoresis of a sample of the amplified clone looked clean, on careful examination of the photograph (Fig 4.8), a slight shadow was visible, supporting the hypothesis that two sequences of similar length had been co-purified.

An attempt to separate these two cDNAs was made by electrophoresing them in an alternative agarose gel, which is designed for the separation of fragments less than 1kb in length (NuSieve GTG agarose, FMC BioProducts, USA). The agarose also has a lower melting temperature, thereby allowing easier extraction of the DNA.

Figure 5.8 clearly shows the two PCR products obtained from the original mixed clone, one product being more abundant than the other. The major product was approximately 872bp and the minor product around 830bp in length.

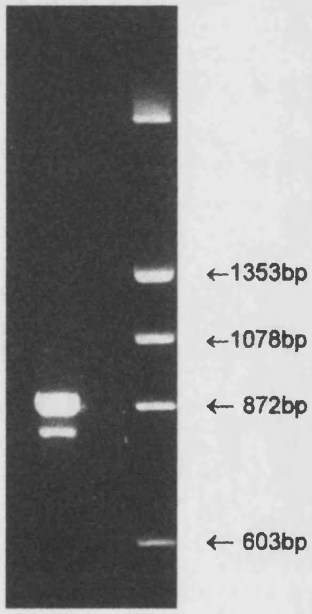


Figure 5.8 Gel electrophoresis in 2% (w/v) NuSieve GTG agarose (FMC BioProducts, USA) of PCR product Z9. Molecular weight markers (M) are the λ *Hin*dIII/ϕX-174 *Hae*III digest (Pharmacia, USA).

Both bands were excised from the gel and the DNA extracted by melting the agarose, as recommended by the manufacturer. Unfortunately this did not work so the samples were extracted using phenol and chloroform in the usual manner. This was also unsuccessful so it was decided to concentrate on other clones which were yielding results and return to these at a later date.

5.2.5 Z11

As seen in Figure 4.8, the amplified PCR product for Z11 was small – approximately 400bp. This enabled the entire sequence to be obtained in one sequencing reaction, using the forward λgt11 primer. The reaction using the reverse primer gave exactly complementary sequence until a section was reached which was AG-rich. At this point, clear, unambiguous sequence with a good strong signal deteriorated into unreadable, small peaks barely above background level, as previously experienced with clone Z5. Once the adapter sequences had been trimmed away, this gave a sequence of 257bp.

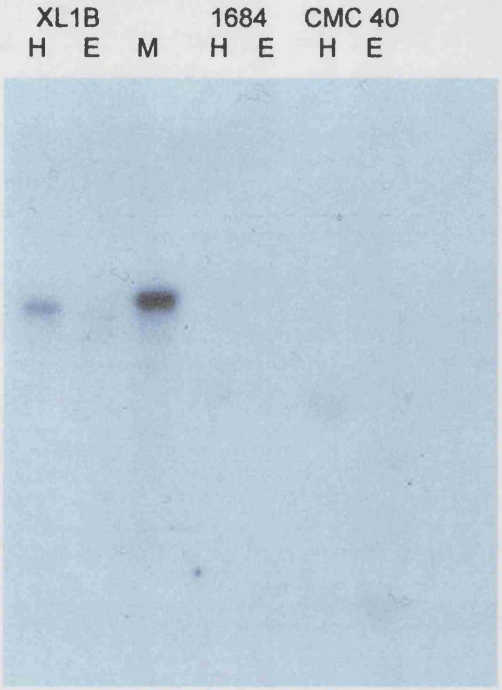
Figure 5.9 shows the results of comparing Z11 with sequences from the GenEMBL database. The first alignment pair show 87% identity with a rat mRNA tandem repeat, with 87% also shown with a different part of the rat sequence moving a single base along the Z11 sequence. The third alignment pair (not shown) was for identical sequence fragments for Z11 (128–97bp and 129–98bp respectively) and two rat genomic sequences, again identical over the lengths in question to 45–35bp and to 2–78bp of Sequence 1.

Figure 5.9 Alignments for Z11 with the following sequences: 1: Rat polymeric Ig receptor tandem repeat mRNA; 2: *Anopheles gambiae* STS DNA.

```
Z11:  128 AAAAGAAGAAAGAAAAGAAGAAGAAGAAGAAA 97
          | ||||||||  | ||||||||||||||||||
Seq 1: 45 AGAAGAAGAAGAAGAAGAAGAAGAAGAAAGAA 35

Z11:  129 GAAAAGAAGAAAGAAAAGAAGAAGAAGAAGAA 98
          || ||||||||  | ||||||||||||||||||
Seq 1:  2 GAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAA 78


Z11:  140 CAGTTGCAGGTGAAAAGAAGAAAGAAAAGAAGAAGAAGAAGAA 98
          || |||  |    |||||||||  | ||||||||||||||||||
Seq 2: 12 CAATTGTTGTACCAAAGAAGAAGAAGAAGAAGAAGAAGAAGAA 78

Z11:  128 AAAAGAAGAAAGAAAAGAAGAAGAAGAAGAAA 9
          | ||||||||  | ||||||||||||||||||
Seq 2: 30 AGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAA 78
```

The location of the rat sequence is in the 3' untranslated region (UTR), between the stop codon and poly (A) tail of the mRNA. The section of Z11 aligning with this region contains repeated sequence based on GAA, where the number of adenosine residues varies from two to four, whereas the rat sequence contains the trinucleotide GAA repeated 10 times. It may be that the cassava repeat represents a microsatellite, or part of one, as they also have short sequence repeats. They are generally not found within the coding parts of genes, so while similarity between a rat Ig receptor and a cassava root sequence would not be expected to indicate a similar function for the two proteins, it may be that the location for the cassava sequence is also in the 3' UTR.

Figure 5.10 Autoradiograph of Southern blot for 10μg CMC 40 and 5μg MCol 1684 total gDNA samples digested singly with *Hin*dIII (H) and *Eco*RI (E), and probed with ^{32}P-labelled PCR product of Z11. XL1B total gDNA was also digested with these enzymes and the marker (M) was the λ *Hin*dIII digest (Pharmacia, USA). Exposure time at –70°C was 3d.



Southern analysis (Figure 5.10) gave a single band in the bacterial sample digested with *Hin*dIII and the marker lane. There is no *Eco*RI site within Z11, so it is strange that it did not hybridize with the second bacterial digest also. This result appeared to indicate that Z11 also was not from cassava.

5.2.6 E28

Clone E28, containing an insert of approximately 850bp in length, was purified from the TAE gel electrophoresis (Figure 4.7) and again sequenced directly with the λgt11 forward and reverse primers. Only the reverse reaction gave a result, with 260bp of readable sequence.

Comparison of this sequence with the databases showed similarity with the *Xenopus laevis* upstream binding factor (UBF) and the nucleolin DNA sequences from

both *X. laevis* (shown here) and rat. A nuclear protein sequence from chicken also showed good similarity and this may in fact be chicken nucleolin. The human nucleolin sequence was aligned with E28 to show similarity at the protein level. When the DNA sequence was translated in all six possible reading frames, homology was still seen with both of these sequences in one of the reading frames (Figure 5.11). It is interesting that in both the alignment with the upstream binding factor and that with nucleolin, the same stretch of E28 was involved.

```
E28: 190 AGGATGTGGTTGAGGATGACGATGATGACGATGATGAGGAAGATGATGATGATGACAAGG
         ||||  ||||||| || || |||||||||||| || || ||||| |||||||| ||| |||
Xl UBF:  AGGATGATGATGACGAAGATGATGATGACGACGAGGATGAAGACGATGATGACGACGAGG

E28:     ATGAAGATGAAGCTGAAGGGCAACAAGATGGAGA 283
         |||||||| | |  ||   |    ||||||||||
Xl UBF:  ATGAAGATAAGGAGGACTCGTCTGAAGATGGAGA 117

E28:     EDDDDDDDEEDDDDDKDEDEAEGQQDGDASGRS 100
         ||| || |||||||| |||      ||| |  |
Xl UBF:  EDDEDDEDEEDDDDDDDEDKEDSSEDGDSSDSS 650


E28:    CTGATAAGCCTGTGGCCGAGGATGTGGTTGAGGATGACGATGATGACGATGATGAGGAAG
        |||     ||||   | ||||||| || |||||| || |||||||| ||||||| || |
Xl nuc: CTGCTGCTCCTGCCTCAGAGGATGAGGATGAGGAAGATGATGATGATGAAGATGATGATG

E28:    ATGATGATGATGACAAGGATGAAGATGAAGCTGA 264
        |||||||||| ||  |||| | || || | |||
Xl nuc: ATGATGATGAAGAGGAGGAGGAGGAAGATGGTGA 854

E28:    DVVEDDDDDDDEEDDDDDKDEDEAEGQQDG---DASGRSK 100
        |  |||||| | |||||  || | |          |  |
Hs nuc: DEDEDDDDDEDDEDDDDEDDEEEEEEEEEEEPVKEAPGKRK 280
```

Figure 5.11 Alignment for clone E28 with part of both the nucleotide and amino acid sequences for the *X. laevis* upstream binding factor (UBF), and with the *X. laevis* nucleolin gene (Xl nuc) and human nucleolin protein (Hs nuc). The UBF nucleotide alignment shows 75.5% identity over 94bp, with the corresponding amino acid homology 63.6%. The percentage homologies for the nucleolin sequences are 73.1% and 52.8% for the nucleotide and amino acid alignments respectively.

Southern analysis (Figure 5.12) gave a complex pattern, with intense hybridization to the four digested cassava gDNA samples. A number of distinct bands can be seen with different mobilities within each of the cassava lanes, with a smear behind them, indicating that E28 is a cassava clone.

The results for E28 were promising, so it was decided to extend the sequence to see whether the homologies above were maintained. Direct sequencing from the PCR products usually gave around 200–300bp. This was enough data for an initial comparison, but as longer runs were routinely available from cloned PCR products, it was thought that sub-cloning would provide a better template for sequencing.

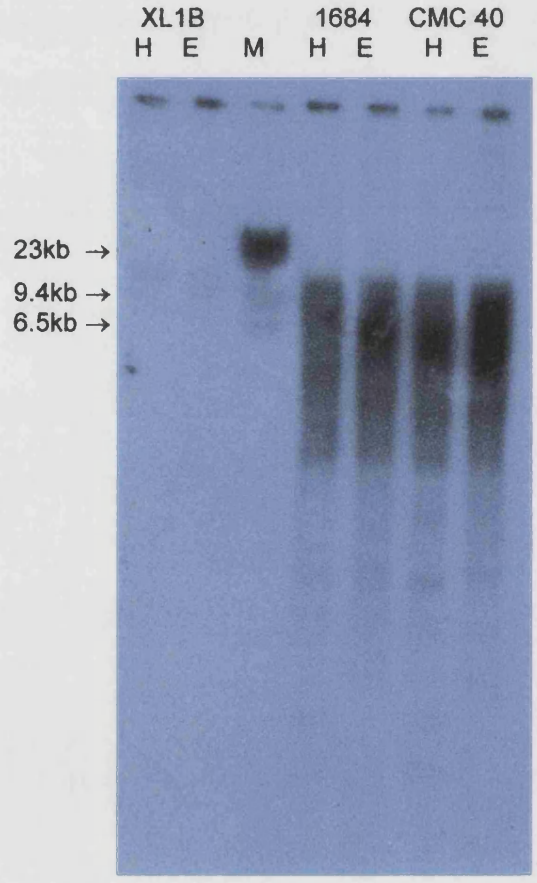


Figure 5.12 Autoradiograph of the Southern blot for 10μg CMC 40 and 5μg MCol 1684 total gDNA samples digested with *Eco*RI (E) and *Hin*dIII (H), and probed with ^{32}P-labelled PCR product of E28. XL1B total gDNA, also digested with these enzymes, and the λ *Hin*dIII digest marker (M). Exposure time at –70°C was 5d.

The vector used for the sub-cloning was pUAg, an updated version of pTAg (R&D Systems, UK). Eleven possible recombinant clones were chosen from the white colonies produced and minipreps made of the DNA. Before sequencing, PCR was carried out using the usual λgt11 primers as a quick check to eliminate non-recombinants. Figure 5.13 shows the amplifications from the 11 putative sub-clones of E28 and a control blue non-recombinant. The amplification product of approximately 850bp was obtained from clones 1–10, but not from clone 11 or the negative control.

Sequencing using the forward and reverse primers on two clones (3 and 5) gave good results. E28 was extended to its full length of 770bp, once the flanking vector sequences had been removed. Figure 5.14 gives the full nucleotide sequence. There were no ambiguities within this sequence and sections covered by multiple fragments were identical in all reactions. Since the longest open reading frame was found in the 3'–5' direction, compared with the first sequencing results from the direct sequencing of the PCR product, the sequence for E28 was reversed.

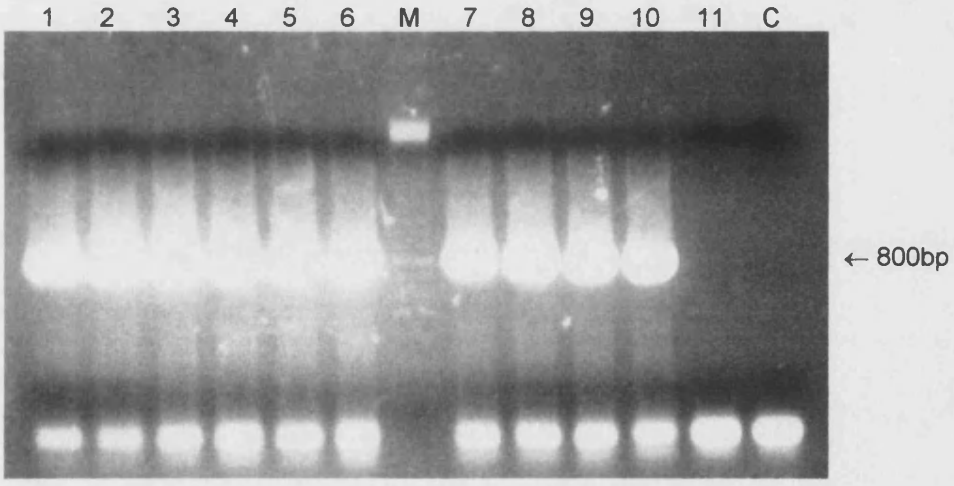


Figure 5.13 Electrophoresis at $5Vcm^{-1}$ in 1% (w/v) agarose of the PCR products from white colonies from the sub-cloning of E28, lanes 1–11. M: 100bp ladder molecular weight markers (Pharmacia, USA) with the 800bp fragment appearing slightly brighter than surrounding bands; C: non-recombinant blue negative control.

Figure 5.14 Full nucleotide sequence of E28, reading in the 3'–5' direction when compared with the original sequencing, for example Figure 5.11. The open reading frame is indicated below the nucleotide sequence. ATG: possible starting methionine residue; TGA: stop codon; TTTGAT: sequence possibly involved with 3' processing.

```
  1  TTAGGGTTTTGCTGTGTTAAGACAGGTACTCGCTCTATTTGTCTATTCTT   50
                          D  R  Y  S  L  Y  L  S  I  L

 51  AATCGAGAAAAAACCATGACTGCCGAGACCCAGGAAGAGCTTGCTGCTCA  100
     N  R  E  K  T  M  T  A  E  T  Q  E  E  L  A  A  Q

101  GCTCGAAGCTCAGAAGATCCTCGAGTCTGATAAGCCTGTGGCCGAGGATG  150
      L  E  A  Q  K  I  L  E  S  D  K  P  V  A  E  D  V

151  TGGTTGAGGATGACGATGATGACGATGATGAGGAAGATGATGATGATGAC  200
       V  E  D  D  D  D  D  D  D  E  E  D  D  D  D  D

201  AAGGATGAAGATGAAGCTGAAGGGCAACAAGATGGAGATGCAAGTGGCAG  250
     K  D  E  D  E  A  E  G  Q  Q  D  G  D  A  S  G  R

251  GTCAAAGCAAAGCAGAAGTGAAAAGAAGAGCAGAAAAGCAATGTTGAAGC  300
      S  K  Q  S  R  S  E  K  K  S  R  K  A  M  L  K  L

301  TTGGAATGAAACCTATCCCTGGTGTTAGTCGTGTCACTGTCAAAAAAAGC  350
       G  M  K  P  I  P  G  V  S  R  V  T  V  K  K  S

351  AAGAATATCTTGTTTGTTATCTCAAAACCTGATGTTTTCAAAAGCCCAAC  400
     K  N  I  L  F  V  I  S  K  P  D  V  F  K  S  P  T

401  TTCAGACACTTATGTTATCTTTGGGGAGGCTAAGATTGAAGACATGAGCT  450
      S  D  T  Y  V  I  F  G  E  A  K  I  E  D  M  S  S

451  CACAACTACAAACTCAGGCTGCAGAGCAGTTCAAGGCTCCTGATCTTAGC  500
       Q  L  Q  T  Q  A  A  E  Q  F  K  A  P  D  L  S

501  CATGTGATTTCCAAGCCTGAGACTTCAACCATGGCTCAAGATGACGAAGA  550
     H  V  I  S  K  P  E  T  S  T  M  A  Q  D  D  E  E

551  GGTGGATGAAACTGGAGTTGAGCCAAAGGATATTGAATTGGTGATGACAC  600
      V  D  E  T  G  V  E  P  K  D  I  E  L  V  M  T  Q

601  AAGCAGGAGTCTCTAGGTCAAAGGCGGTCAAGGCTTTGAAGGCTGCTGAT  650
       A  G  V  S  R  S  K  A  V  K  A  L  K  A  A  D

651  GGAGACATTGTTTCTGCCATAATGGAGCTCACCAATTGATAGGAGCCTTG  700
     G  D  I  V  S  A  I  M  E  L  T  N  *

701  CAGAAACTGTTATGTGATCTGGTGGCATTATTATTATGGTTCTTCATATT  750

751  GTTTGATCTGAAATCAGGAATTTTTTGTC                       779
```

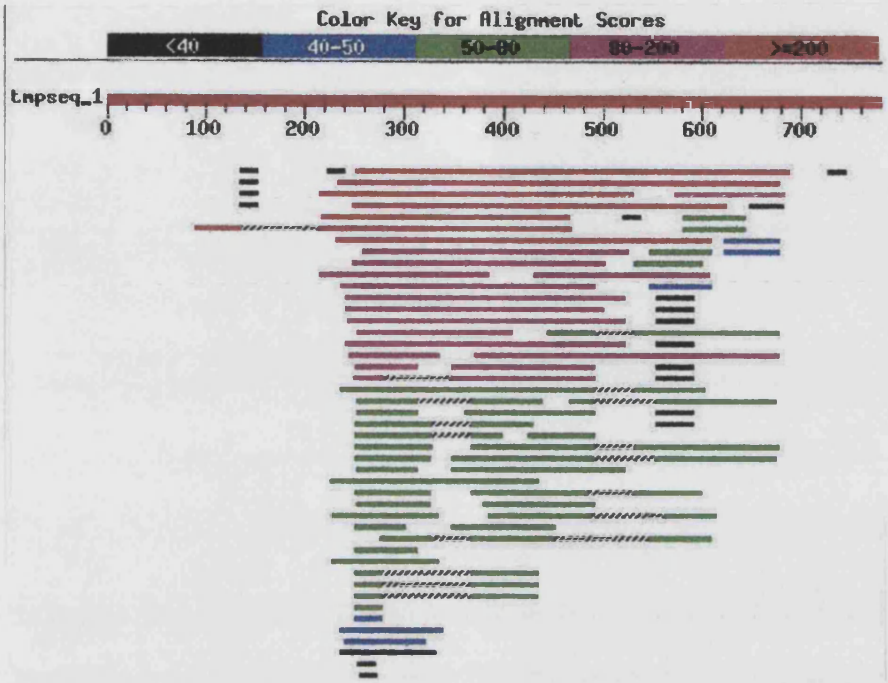
When this extended sequence was compared again with the GenEMBL database, homology was seen with the nucleotide sequence from an expressed sequence tag (EST) from rice. Over 369bp there was 78.9% identity. This actually increased at the translated amino acid level, with the same area giving 89.5% identity in 123 amino acid overlap (seen in Figure 5.15).

An advanced BLAST search (Altschul *et al.*, 1997) gave comparisons for E28 with expressed sequences from a wide range of plant tissues, including woody (aspen) and non-woody (cotton), specialized (ovary, endosperm) and undifferentiated (callus) tissues (Fig. 5.16). The two best nucleotide alignments from these results are shown in Figure 5.17. The second alignment gives the first indication of a function for E28 – similarity with a cDNA from cotton, which is itself similar to the nascent polypeptide associated complex and activator α chain (α-NAC) mRNA from *A. thaliana*.

```
     70          80          90         100         110         120
E28:  QDGDASGRSKQSRSEKKSRKAMLKLGMKPIPGVSRVTVKKSKNILFVISKP
        |||||||||||||||||||| ||||||:|:|||||:||||||||||||||
EST:    DASGRSKQSRSEKKSRKAMQKLGMKTITGVSRVTIKKSKNILFVISKP

               130         140         150         160         170
E28:   DVFKSPTSDTYVIFGEAKIEDMSSQLQTQAAEQFKAPDLSHVISKPETST
       ||||||:|||||||||||||:|||||||||||||||||:||||:|:|:
EST:   DVFKSPNSDTYVIFGEAKIEDLSSQLQTQAAEQFKAPDLSNVISKAEPSA

               180         190
E28:   MAQDDEEVDETGVEPKDIELVMTQA
        |||||||||:||||||||||||||:
EST:   AAQDDEEVDESGVEPKDIELVMTQG
```

Figure 5.15 E28 amino acid alignment with that from a rice EST. Numbers refer to the position of the amino acid below the first digit within the E28 translated sequence seen in Figure 5.14. Identical residues are shown by dashes; residues whose properties are similar but are not identical are shown by dots.

Figure 5.16 Results from the BLAST search for E28, listing the first 30 alignments. The first two alignments, with E28 compared with the aspen and cotton sequences, are shown in full (Figure 5.17).



gb|AI166170|AI166170_A013P70U Hybrid aspen plasmid library...	446	e-124
gb|AI726778.1|AI726778 BNLGHi6564 Six-day Cotton fiber *Goss*...	260	1e-67
gb|AI731383.1|AI731383 BNLGHi8891 Six-day Cotton fiber *Goss*...	246	1e-63
gb|AI165642|AI165642 A087p59u Hybrid aspen plasmid library ...	230	9e-59
gb|AI165249|AI165249 A079P49U Hybrid aspen plasmid library ...	220	8e-56
gb|AI162667|AI162667 A021P36U Hybrid aspen plasmid library ...	208	3e-52
gb|AI486280.1|AI486280 EST244601 tomato ovary, TAMU *Lycoper*...	208	3e-52
gb|AI899415.1|AI899415 EST268858 tomato susceptible, *Cornel*...	194	5e-48
gb|AI441868|AI441868 sa53d09.y1 Gm-c1004 *Glycine max* cDNA c...	157	1e-36
gb|AI745926.1|AI745926 605076B10.x1 605 - Endosperm cDNA li...	127	9e-28
gb|AI161706|AI161706 A005P55U Hybrid aspen plasmid library ...	125	4e-27
gb|AI164717|AI164717 A067P60U Hybrid aspen plasmid library ...	123	1e-26
gb|AI099855|AI099855 34008 Lambda-PRL2 *Arabidopsis thaliana*..	103	1e-20
gb|AI665337.2|AI665337 605010A05.x1 605 - Endosperm cDNA li...	103	1e-20
gb|AI795478.1|AI795478 605010A05.y1 605 - Endosperm cDNA li...	101	5e-20
gb|AI737778.1|AI737778 606041B06.x1 606 - Ear tissue cDNA l...	101	5e-20
gb|AI165465|AI165465 A084P45U Hybrid aspen plasmid library ...	100	2e-19
gb|AI637381.1|AI637381 DB# 320 Giant cell specific cDNA *Lyc*...	100	2e-19
gb|AI897368.1|AI897368 EST266811 tomato ovary, TAMU *Lycoper*...	98	8e-19
gb|AI820387.1|AI820387 605092H07.y1 605 - Endosperm cDNA li...	96	3e-18
gb|AI779799.1|AI779799 EST260678 tomato susceptible, *Cornel*...	94	1e-17
gb|L46497|BNAF1979 *Brassica campestris* (clone F1979) expres...	88	8e-16
gb|AI441117|AI441117 sa59a07.y1 Gm-c1004 *Glycine max* cDNA c...	86	3e-15
gb|AA143915|AA143915 zEST00737 Maize Leaf, Stratagene #9370...	82	5e-14
gb|AI665553.1|AI665553 605012G07.x1 605 - Endosperm cDNA li...	82	5e-14
gb|AI484194.1|AI484194 EST249347 tomato resistant, Cornell ...	80	2e-13
dbj|D15896.1|D15896 RICC1482A Rice callus *Oryza sativa* cDNA...	80	2e-13
b|AI780994.1|AI780994 EST261873 tomato susceptible, *Cornel*...	80	2e-13
gb|AI670210.1|AI670210 605019H08.x2 605 - Endosperm cDNA li...	78	8e-13
gb|AI441575|AI441575 sa62b06.y1 Gm-c1004 *Glycine max* cDNA c...	76	3e-12

Figure 5.17 Two best nucleotide alignments from BLAST search shown in Figure 5.16. Alignment 1 shows E28 compared with a cDNA sequence for aspen (A).

Alignment 1:

```
E28: 252 tcaaagcaaagcagaagtgaaaagaagagcagaaaagcaatgttgaagcttggaatg 311
         ||||| ||||||| |||||||||||||| ||  | ||||||||||||||||||||||||
A:    13 tcaaaacaaagccgaagtgaaaagaaaagtcgtaaagcaatgttgaagcttggaatg 72

E28: 312 aaacctatccctggtgttagtcgtgtcactgtcaaaaaaagcaagaatatcttgttt 371
         ||| ||||||||||||||||||| || || || ||||| ||||||||||||||||||
A:   73  aaaactatccctggtgttagtcgggttacagtaaaaaagagcaagaatatcttgttt 132

E28: 372 gttatctcaaaacctgatgttttcaaaagcccaacttcagacacttatgttatcttt 431
         |||||||| |||||||||||||||| |||||||| || ||||| ||||| ||||||
A:   133 gttatctccaaacctgatgttttcaagagcccaacatctgacacatatgtaatcttt 192

E28: 432 ggggaggctaagattgaagacatgagctcacaactacaaactcaggctgcagagcag 491
           |||||||||||| |||||| |||| || || ||||||||||||| |||||||||
A:   193 ggagaggctaagatcgaagacctgagttcccagctacaaactcaggcagcagagcag 252

E28: 492 ttcaaggctcctgatcttagccatgtgatttccaagcctgagacttcaaccatggct 551
         |||||||||||||||| || ||||| ||||| || |||||||||||| |||| |||
A:   253 ttcaaggctcctgatctcagtcatgtcatttcaaaacctgagacttcagccatagct 312

E28: 552 caagatgacgaagaggtggatgaaactggagttgagccaaaggatattgaattggtg 611
         || ||||| |||||||| |||||||| ||||| |||||||||||||||||||| |||
A:   313 caggatgatgaagaggtagatgaaaccggagtggagccaaaggatattgaattagtg 372

E28: 612 atgacacaagcaggagtctctaggtcaaaggcggtcaaggctttgaaggctgctgat 671
         ||||| || |||||||||||| |||||||||| || || |||||||||||| || ||
A:   373 atgacccaggcaggagtctccaggtcaaaggctgtgaaagctttgaaggcagcagat 432

E28: 672 ggagacattgtttctgccataatggagctcaccaattga 689
         |||||||| |||||||||||||||||||| |||||||||
A:   433 ggagacatagtttctgccattatggagcttaccaattga 450
```

Figure 5.17 (continued) Alignment 2 shows E28 compared with a cDNA sequence from cotton similar to the nascent associated complex alpha chain from *Arabidopsis thaliana* (B).

Alignment 2:

```
E28: 234 ggagatgcaagtggcaggtcaaagcaaagcagaagtgaaaagaagagcagaaaagca 293
         ||||||| || ||| ||||||||||||||| ||||||||||||||||  | |||||
B:   180 ggagatggaactggtaggtcaaagcaaagccgaagtgaaaagaagagtcgcaaagct 239

E28: 294 atgttgaagcttggaatgaaacctatccctggtgttagtcgtgtcactgtcaaaaaa 353
         |||||||||||||| |||||||| |||||||||||||| || |||||||| || ||
B:   240 atgttgaagcttgggatgaaaccgatccctggtgttagccgggtcactgtgaagaag 299

E28: 354 agcaagaatatcttgtttgttatctcaaaacctgatgttttcaaaagcccaacttca 413
         |||||||||||| || ||||| |||||||||| || || ||||| |||||| | |||
B:   300 agcaagaatattttatttgtcatctcaaaaccagacgtgttcaagagcccagcatca 359

E28: 414 gacacttatgttatctttggggaggctaagattgaagacatgagctcacaactacaa 473
         || |||||||| || ||||| |||||||||||||| ||| | |||||||||||| ||
B:   360 gatacttatgtaatatttggagaggctaagattgaggacttaagctcacaactgcag 419

E28: 474 actcaggctgcagagcagttcaaggctcctgatcttagccatgtgatttccaagcct 533
           ||||  |||||||||||| |||||||||||||||| ||||| ||||| || ||
B:   420 actcgagctgcagagcaattcaaggctcctgatctaagccaagtgatatcaaaacca 479

E28: 534 gagacttcaaccatggctcaagatgacgaagaggtggatgaaactggagttgagcca 593
         ||| | ||||| |  || || || ||||| |||||||| || || || |||||
B:   480 gagtcatcaacagcagtccaggacgatgaagaagtggatganacaggggtggagccc 539

E28: 594 aaggatattgaattggtgatgacacaagcaggagtctctaggtcaaaggcggtcaag 653
         ||||| |||||||| ||||||||||||||||||| || | ||  ||||| ||||||
B:   540 aaggacattgaattatttatgacacaagcaggagtatccangttcaaggctgtcaag 599

E28: 654 gctttgaaggctgctgatggagacattgtttctgccataatggagct 679
         ||  | |||||||||||||| |||||||||  ||| ||||||||||
B:   600 gcactcaaggctgctgatggtgacattgttattgctataatggagct 625
```

A truncated version of E28, comprising amino acids 75–207, was used in another BLAST search as the sequences most similar to the full version were EST sequences, which did not confirm the function. This smaller fragment was 73% homologous with a section of the putative α-NAC sequence from *A. thaliana*. The full *Arabidopsis* sequence was then used in an alignment with E28, as shown in Figure 5.18. This is a possible alignment, since the two proteins differ in size, with E28 being 207 amino acids long while the *Arabidopsis* protein is 212.

Figure 5.18 Alignment for translated E28 with a putative α-NAC from *A. thaliana* (NAC). Results are from the BLAST search using a truncated version of E28 (75–207 amino acids). Identical residues are indicated by the appropriate letter while those residues that are not identical but have similar properties are indicated by + .

```
E28:   1 MTAETQEELAAQLEAQKILESDKPVAE-DVV-EDDDDDDDEEDDDDDKDE  48
         M     EE          I E  K   E DVV ED  D D+++DD DD D+
NAC:   1 MPGPVIEEVNEEALMDAIKEQMKLQKENDVVVEDVKDGDEDDDDVDDDDD  50

E28:  49 DEAEGQQDGDASGRSKQSRSEKKSRKAMLKLGMKPIPGVSRVTVKKSKNI  97
         + A G     A   SKQSRSEKKSRKAMLKLGMKP+  VSRVT+K+SKN+
NAC:  51 EIADGAGENEA---SKQSRSEKKSRKAMLKLGMKPVTDVSRVTIKRSKNV  96

E28:  99 LFVISKPDVFKSPTSDTYVIFGEAKIEDMSSQLQTQAAEQFKAPDLSHVI 148
         LFVISKPDVFKSP S+TYVIFGEAKI+DMSSQLQ QAA++FK PD++ +I
NAC:  98 LFVISKPDVFKSPNSETYVIFGEAKIDDMSSQLQAQAAQRFKMPDVASMI 147

E28: 149 SKP---ETSTMAQ---DDEEVDETGVEPKDIELVMTQAGVSRSKAVKALK 192
               E +T+AQ   DDE+VDETGVE KDI+LVMTQAGVSR KAVKALK
NAC: 148 PNTDGSEAATVAQEEEDDEDVDETGVEAKDIDLVMTQAGVSRPKAVKALK 197

E28: 193 AADGDIVSAIMELTN  207
          ++GDIVSAIMELT
NAC: 198 ESNGDIVSAIMELTT  212
```

5.3 DISCUSSION

5.3.1 The 'A' and 'Z' clones

Of the six DNA sequences investigated in this chapter, A35 and A45 were apparently bacterial in origin. This was disappointing, but not surprising as storage roots naturally contain bacteria. For example, it has been reported in potato tubers that 25 bacterial species from 18 genera represented 73% of the total bacterial content (Sturz *et al.*, 1998).

The Z clones yielded little useful information. Z5 and Z11 were both A-rich, resulting in difficulty with obtaining sequence data, while the sequencing of Z9 revealed

it was a mixed product containing two cDNAs. The BLAST and FASTA searches of the database (Pearson and Lipman, 1988; Altschul *et al.*, 1990, 1997) gave limited information as to possible identities for Z5 and Z11. The latter contained an AAG tandem repeat, which may indicate that this section of Z11 represents a part of the 3' UTR. Sub-cloning these fragments would probably improve the sequencing and give more data for comparison.

Southern analysis was only obtained for Z11, and appeared to indicate that it was not a cassava clone, although this would need to be repeated since hybridization only occurred with one of the bacterial samples. It is possible that the presence of a repeated sequence that may be a microsatellite in Z11 distorted the result.

5.3.2 Clone E28 – a cassava α-NAC homologue

The most interesting clone was E28, shown to be a genuine cassava cDNA. Within the full cDNA sequence was a section containing 369bp (123 amino acids) which was homologous with an expressed sequence from the rice genome at both the DNA and protein levels. This gave the best homology, but did not allow identification of E28. Further BLAST searches using the non-human non-mouse EST databases gave alignments with a wide range of sequences, mainly from plants, the best of which was 88.8% homology over 437bp with a *Populus* sequence. By translating this portion of the sequence, 73% similarity at the protein level was obtained over this truncated length of 132 amino acids for α-NAC from *Populus* and *A. thaliana* respectively. α-NAC is a protein found across many species from *S. cerevisiae* to humans, is involved in binding the nascent RNA emerging from the ribosome and thought to be involved in signal-sequence-specific protein sorting (Wiedmann *et al.*, 1994).

The yeast homologue, EGD2, is involved in carbohydrate metabolism and may act as a transcription factor exerting a negative effect on the expression of several genes transcribed by RNA Polymerase II (Reimann *et al.*, 1999). The early BLAST

results using the original fragment of E28 showed homology over 93bp of the nucleotide sequence and an aspartic acid-rich stretch from amino acid residues 65–100. These similarities were not with α-NAC/EGD2, but with an UBF and nucleolin. UBF is a transcription factor, involved with initiation of transcription by RNA polymerase I (Learned *et al.*, 1986), while nucleolin is an RNA-binding protein (Ghisolfi-Nieto *et al.*, 1996). In nucleolin, this acidic region represents one of the RNA-binding regions.

When the protein sequence for E28 was compared with the full protein sequence for *A. thaliana* putative α-NAC, the identity was found to extend 5' to that already shown for the truncated version for another 12 residues i.e. amino acids 74–63 of E28 (Fig. 5.18). The two sequences then diverge, but that for *A. thaliana* also contains an aspartic acid-rich domain, located at amino acids 37–50, supporting the suggestion that this domain is involved with RNA binding. Both proteins are a similar length; 212 and 207 amino acids for *A. thaliana* and cassava respectively.

A final observation was the Southern results, which indicated that E28 was part of a multigene family. The EGD2/α-NAC genes have also been shown to be part of such families (Shi *et al.*, 1995). Thus the cassava sequence E28 appears to be an α-NAC homologue, showing good similarity with other α-NACs over its carboxy end and having homology with known RNA-binding domains towards the N-terminus.

5.3.2.1 3' processing elements

The region 3' to the stop codon at 687–689bp was examined to see whether it contained any of the motifs relating to polyadenylation that have been found in other plant genes and their respective cDNAs. The polyadenylation signal AAUAAA, found in animal and yeast sequences, is less conserved in plants and is missing entirely from some genes (Heidecker and Messing, 1986; Hughes, 1996). This is located 10–40 nucleotides upstream of the cleavage site of the mRNA and is referred to as the near upstream element (NUE). Moving further 5' from the cleavage site is an area

containing the far upstream element (FUE), which appears to be a general requirement for 3' processing (Rothnie, 1996). Although there does not seem to be a consensus sequence, FUEs are U- or UG-rich and the hexanucleotide UUUGUA, or a derivative thereof, is represented one or more times in many FUEs. A possible candidate in E28 is the element TTTGAT within the cDNA sequence, at 751–756bp (underlined in Figure 5.14), indicating that the NUE and polyadenylation site had not been reached before the available sequence ended. Comparing the distances involved with the cDNA sequence for cassava granule-bound starch synthase (Salehuzzamen *et al.*, 1993), the latter shows a distance of 255bp between the stop codon and the poly (A) tail, with the AATAAA signal 20bp upstream of the cleavage site.

Further work on E28 would be to determine the tissue expression by northern analysis. From the diversity of tissues from which α-NAC sequences have been derived, it would be expected that it would also be expressed in all tissues of cassava. Other experiments could be carried out to see whether all copies of the cassava gene are expressed or whether some have lost functionality.

CHAPTER 6: PHYLOGENETIC ANALYSIS OF CASSAVA USING rRNA SEQUENCES

6.1 INTRODUCTION

6.1.1 Traditional classification

Taxonomy, or the study of the classification of biological organisms, has been of great interest to biologists for centuries. One early attempt to organize the living world into groups was that of Aristotle (384–322 BC), based on whether an organism possessed red blood or not. More recently, in 1759, Carolus Linnaeus introduced the hierarchical system whereby groups of organisms are further subdivided until the individual itself is reached. This system, together with Linnaeus' binomial nomenclature, persists today as the means by which many organisms are classified.

Traditionally, classification relied on the comparison of morphological characteristics. However, this often lacked reliable parameters for meaningful comparisons and was therefore highly subjective. Another problem was that it required great experience and familiarity with the taxonomic group in question, so was limited to a few specialists. These factors made it difficult to challenge proposed classifications.

6.1.2 Evolutionary taxonomy

With the publication of Darwin's *On the Origin of the Species ...* (1859), the concept of evolution added another facet to classification. The question of where an organism should be placed in relation to other species, and possible lines of evolutionary descent, became issues to which biologists sought answers. Thus results from classification studies were required to do more than simply group individuals

together, they were expected to provide information on an organism's phylogeny, or evolutionary history.

6.1.3 Phenetics versus cladistics

The approach of phenetics, or numerical taxonomy, was introduced to address the issues of subjectivity and the lack of uniform measures for comparison mentioned above. Phenetics groups organisms by overall similarity using as many measurable traits as possible. This approach was criticized in turn, particularly by cladists, who claimed that phylogenetic relationships cannot be deduced by overall similarity but are reflected in the similarity attributable to those derived characters shared by two or more descendent taxa.

For example, one of the main objections to phenetics is that it gives little information regarding the evolutionary lineages between different taxonomic groups. Thus, while classification based on a set of characters such as swim bladder position and gill rakers for fish, and plumage feathers and toes on the feet of birds, are useful within each group they do not allow easy comparison between the groups. There is also the question as to whether each character is given equal weighting, since organisms can be compared even when a character is absent provided there is reasonable overall similarity. However, calls for higher weighting of certain characters are countered by claims of a loss of objectivity.

A second issue is that of homoplasy, or convergent evolution. Based on phenetics, organisms which have independently evolved particular characteristics can be grouped closer than they should. The classic example here is that of vultures from the 'Old' and 'New' worlds, which both show adaptations such as the bald head, powerful hooked bill and soaring behaviour for feeding on carrion. Whilst originally grouped together taxonomically by phenetics, molecular analysis has placed the 'New' world vultures closer to storks.

Cladistics is not without its critics. Particularly, the concern that a single shared-derived trait is sufficient to show a unique evolutionary relationship prompts the call for large numbers of characters to be assayed. However, cladistics does attempt to distinguish between ancestral and shared-derived traits, the most common method of determining that a character is primitive being its presence in an outgroup.

Thus the two kinds of data available for phylogenetic analyses are distances and discrete characters. (For a review see Avise, 1994.)

6.1.4 Molecular data

During the second half of the twentieth century, there was a change in the approach towards phylogenetics. The use of molecular data, firstly from proteins and then from DNA, became widespread. The advantage of using these data is that it is known that they are genetic and therefore represent consequences of evolution. With the discovery of restriction enzymes, or endonucleases, which cut DNA at sequence-specific sites (Linn and Arber, 1968; Meselson and Yuan, 1968) and the techniques of DNA-DNA hybridization (Southern, 1975) and DNA sequencing (Maxam and Gilbert, 1977; Sanger *et al.*, 1977), the possibility to compare the actual DNA sequences from different organisms across separate kingdoms also became a reality.

During the 1970s, a major contribution to systematics by molecular data was the classification of the Archaea. Prior to this, living organisms had been divided phenotypically into eukaryotes, or those possessing a true cell nucleus, and prokaryotes, which do not. Using the slowly evolving 16S ribosomal DNA sequences as data, Woese and colleagues reorganized this classification into three major domains; Eukarya, Eubacteria and Archaebacteria (later called Archaea) (Fox *et al.*, 1977; Woese *et al.*, 1990). This showed that the prokaryotes were not a single entity and furthermore that the Archaebacteria were actually more closely related to the Eukarya than the Eubacteria. Since then, comparison of DNA and protein sequences

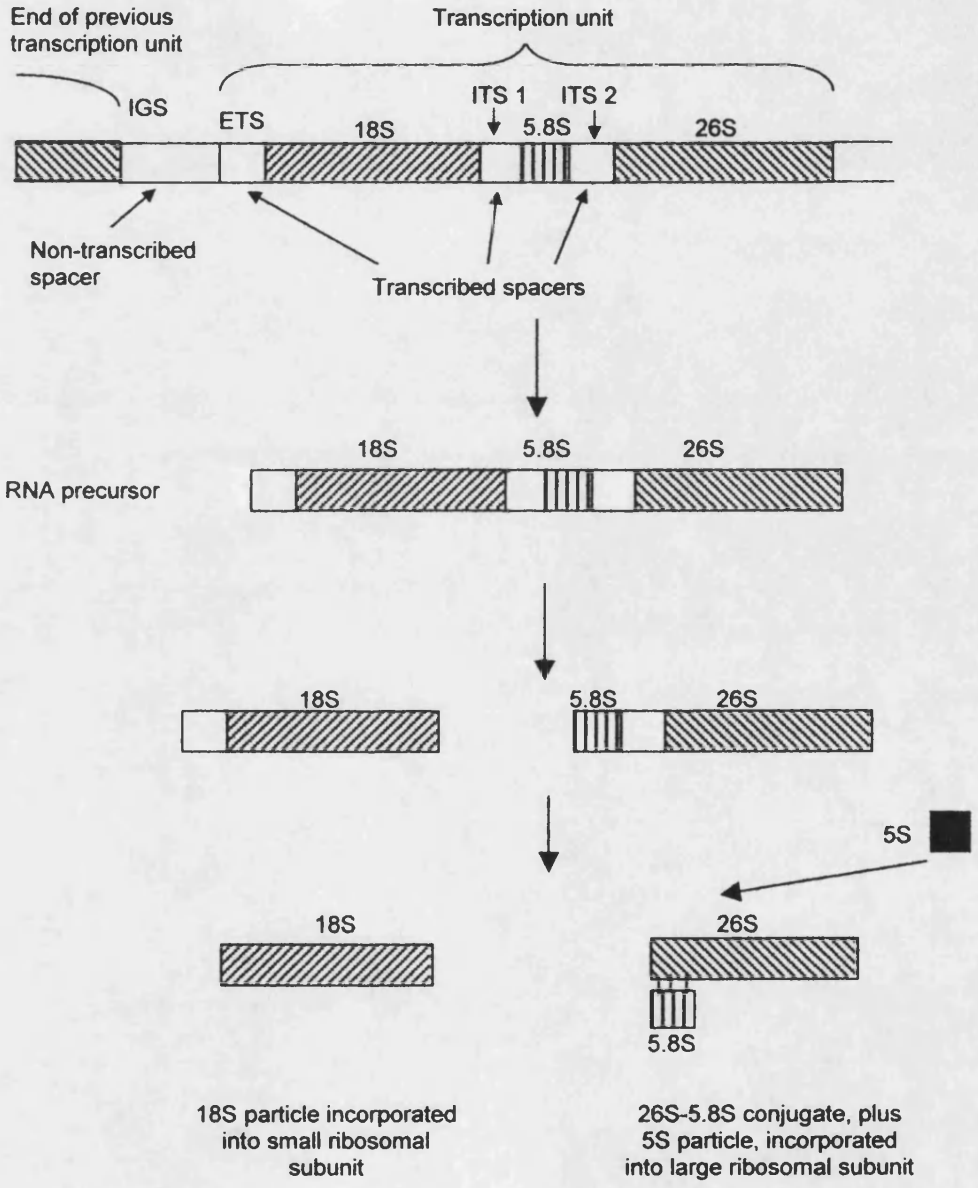
has become commonplace, and with numerous genome sequencing projects underway the amount of data available is increasing daily.

6.1.5 Use of rRNA gene data

Ribosomal RNA genes have been used extensively to investigate phylogenetic relationships. These genes code for the functional parts of the ribosome, which bind to mRNA in the presence of various initiation factors, and are responsible for protein synthesis. Present in the nuclei of prokaryotes and eukaryotes, the mitochondria of eukaryotes and the chloroplasts of plants, rRNA genes are composed of highly conserved coding regions interspersed with more variable non-coding spacer regions (see Figure 6.1). In the nucleus they are generally arranged as tandem repeats, with a highly variable copy number. For example, *E. coli* has seven, while humans have around 200 and plants can have several thousand (overview, Alberts *et al.*, 1994).

For individuals, and within a species, rRNA genes show a higher level of sequence similarity than would be expected if all parts of the genome were evolving at the same rate, with both coding and internal spacer regions virtually identical for the numerous repeat units. Between different species, the exons are still conserved, but the intergenic regions are highly divergent. This intraspecific homogeneity and interspecific heterogeneity is referred to as concerted evolution, and is thought to arise from unequal crossover and gene conversion (Hamby and Zimmer, 1992). The length of the intergenic spacers (IGS) can vary considerably, within individuals and even at their different chromosomal loci, due to varying numbers of subrepeats in the middle IGS region. Copy number in plants is also highly variable, with different tissues even having different numbers of repeat units.

Figure 6.1 Representation of the transcription unit of plant nuclear rRNA and its processing into the large and small ribosomal subunits. Coding regions are represented by the shading (▨ ,◫ and▧). Non-coding spacers are blank (□) are of variable length and consist of non-transcribed spacers and transcribed spacers. The latter are named according to their location within the transcription unit as external transcribed spacers (ETS) and internal transcribed spacers (ITS). Overall length of a single plant transcription unit is 11–13kb (after Hamby and Zimmer, 1992; Alberts *et al.*, 1994).



Different parts of the repeat unit have been used in different phylogenetic analyses (for review see e.g. Avise, 1994). For example, the nuclear small subunit sequences which evolve slowly have been used to study relationships between distantly related organisms, such as that of the Eubacteria and Archaea mentioned above. In plants, Hamby and Zimmer (1992) used nuclear large and small rRNA sequences to analyse relationships between 60 diverse taxa and suggested relationships regarding angiosperms and gymnosperms, and monocots and dicots.

Data from the more rapidly evolving mitochondrial rRNA genes have also been used to analyse more recent evolutionary relationships. For example, the hypothesis that mammalian flight, as exhibited by bats, evolved only once was investigated using DNA sequences from the 12S rRNA gene, together with the mitochondrial cytochrome oxidase gene. Traditionally megabats ('flying foxes') and microbats (typical bats) are considered the closest relatives to each other i.e. divergence of primates preceded the evolution of flight, with microbats and megabats diverging later. Based on newly observed neuroanatomical features, an alternative hypothesis was proposed. Here, microbats diverged first, then evolved flight, with the split between primates and megabats occurring later, to be followed by a second evolution of flight along the megabats branch. Results from the molecular analysis gave strong support for the evolution of flight only once (review, Avise, 1994).

Finally, the non-coding regions of the rRNA transcription unit have been used, both the internal transcribed spacers (ITS) and the intergenic spacers (see Figure 6.1). These are much more rapidly evolving than any of the coding regions, and have been used to compare different populations, or members within a genus or family. An example is that of the genus Phytophthora, where four different species (*P. palmivora*, *P. megakarya*, *P. capsici* and *P. citrophthora*) thought to be closely related were analysed together with a fifth Phtyophthora species (*P. cinnamomi*) used as an outgroup (Lee and Taylor, 1992). A common lineage was suggested from the

molecular results for the first two species, and a close relationship between *P. capsici* and *P. citrophthora*, while large distance values were obtained between these four species and *P. cinnamomi*.

These results show the use of different parts of the rRNA transcription units over a wide range of species in the analysis of molecular data for inferring phylogenies.

6.1.6 Phylogenetic algorithms

Many phylogenetic algorithms exist and which method to use to analyse data is an issue of debate. Distance methods convert the aligned sequences to a pairwise matrix, then build the evolutionary tree, whereas methods using discrete character states consider each nucleotide, or the function of each site, directly. Examples of these two methods are neighbour-joining and maximum parsimony, respectively, and a number of computer programmes exist to facilitate these analyses. Neighbour-joining is a 'cluster' method, i.e. an algorithm sets up a tree for three sequences then adds the remainder one by one. This method is quick and easy but the result is often dependent on the order in which the sequences are added. Additionally it cannot evaluate two different hypotheses. Search methods, such as maximum parsimony, choose among the set of all possible trees with a score assigned to each tree based on some measure of how the data and tree relate to each other. The most parsimonious tree is that requiring the least number of evolutionary changes to explain the differences observed between the operational taxonomic units used in its construction, i.e. to fit the data. Search methods have the advantages over distance methods of allowing the evaluation of the quality of any one tree, and of requiring a function to relate both data and tree. However, they are time-consuming and for any reasonable number of sequences it is often impossible to guarantee that the optimal tree has been found. Thus the different methods all have certain limitations and these should be borne in mind when interpreting the results. (For a review, see e.g. Page

and Holmes, 1998.)

Of the clones isolated from the cassava cDNA library in Chapter 4, three are discussed here. A24, A30 and E30 were investigated further and after the usual procedures of cloning, sequencing and Southern analysis found to represent partial ribosomal sequences from cassava. These were subsequently used in phylogenetic analyses.

6.2 MATERIALS AND METHODS

6.2.1 Sub-cloning

Both A24 and A30 were sub-cloned into pTAg, as described for A35 in Chapter 5. Plate lysates were made to obtain template DNA for PCR, then the usual large-scale PCR amplification was carried out using the λgt11 forward and reverse primers. The fragments were purified by excising the bands containing the PCR products from a preparative gel and extracting the DNA by the Sephaglas method. After checking the quality of the DNA on an agarose gel and determining the concentration, 50ng A24 and 20ng A30 were used for cloning into the vector.

E30 was cloned into the pGEM-T Easy vector (Promega, USA), using two different insert:vector ratios 1:1 and 3:1 (i.e. 13.33ng and 40ng E30 DNA respectively).

6.2.2 Sequencing

Manual sequencing in both directions was used to obtain data from three clones for each of A24 and A30, using the λgt11 forward and reverse primers. Primers located within the pTAg vector were also used, reading in both 5' and 3' directions.

The PCR product of E30 was electrophoresed, excised and purified before being sequenced directly using the λgt11 reverse primer. E30 was then sub-cloned into the vector pGEM-T Easy (Promega, USA). As usual, recombinants were selected

by single-colony gel analysis and DNA prepared by Wizard miniprep method. Using the pGEM-T Easy primers in addition to those for λgt11, complete sequence was obtained in both directions for E30.

6.2.3 Preparation of further DNA

More DNA was required for A24 and A30, prepared here by transforming TSS competent cells. After transformation, pTAg-A24 and pTAg-A30 were plated out on agar plates supplemented with kanamycin and single colonies grown overnight in a mini-culture of 2TY supplemented with ampicillin, thereby providing a double selection procedure. DNA was prepared by the Wizard method from colonies that grew successfully, and the quality verified as usual.

6.3 RESULTS

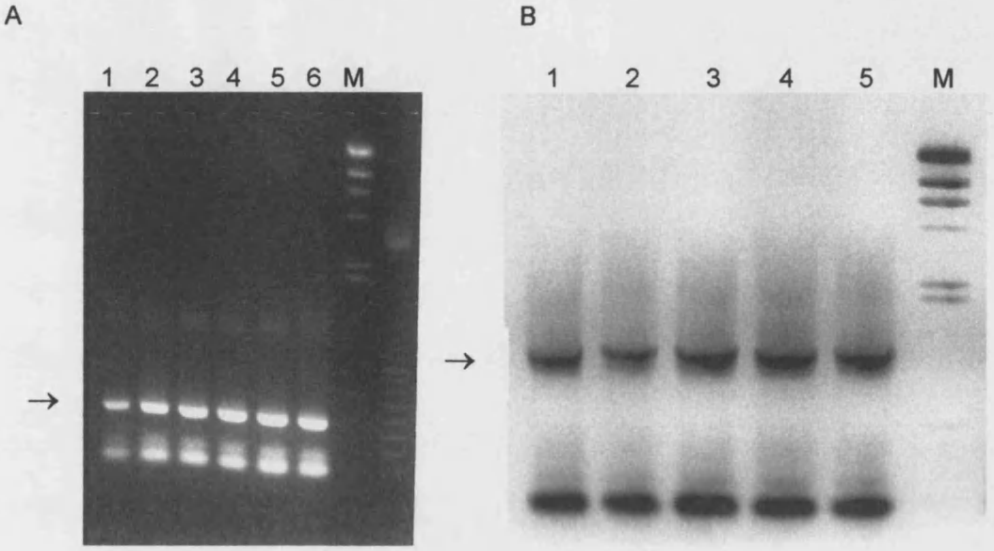
6.3.1 Identification of clones for the cassava 26S ribosomal subunit

6.3.1.1 Sub-cloning

Figure 6.2 shows the results for the amplifications of the two cloned fragments A24 and A30, with products of the correct sizes obtained. A24 was 350bp while A30 was 1.2kb, the bands indicated by arrows.

As usual a number of the white colonies were selected for single-colony gel analysis, together with colonies containing the self-ligated plasmid and the control insert. From these results, (Figure 6.3), DNA was prepared from those which appeared to be genuine recombinants and confirmed by checking samples using gel electrophoresis.

Figure 6.2 Electrophoresis in 1% (w/v) agarose at $5Vcm^{-1}$ for the PCR products (A) A24 and (B) A30, both amplified using λgt11 primers. Products of the expected sizes of 350bp for A24 (A; lanes 1–6) and 1.2kb for A30 (B; lanes 1–5) can be seen, indicated by the respective arrows. Markers (M) were λ DNA digested with *Hin*dIII in both cases. Since excess primers were used to avoid this limiting the reaction, these are seen as low molecular weight bands below the PCR products. B is a reverse picture.



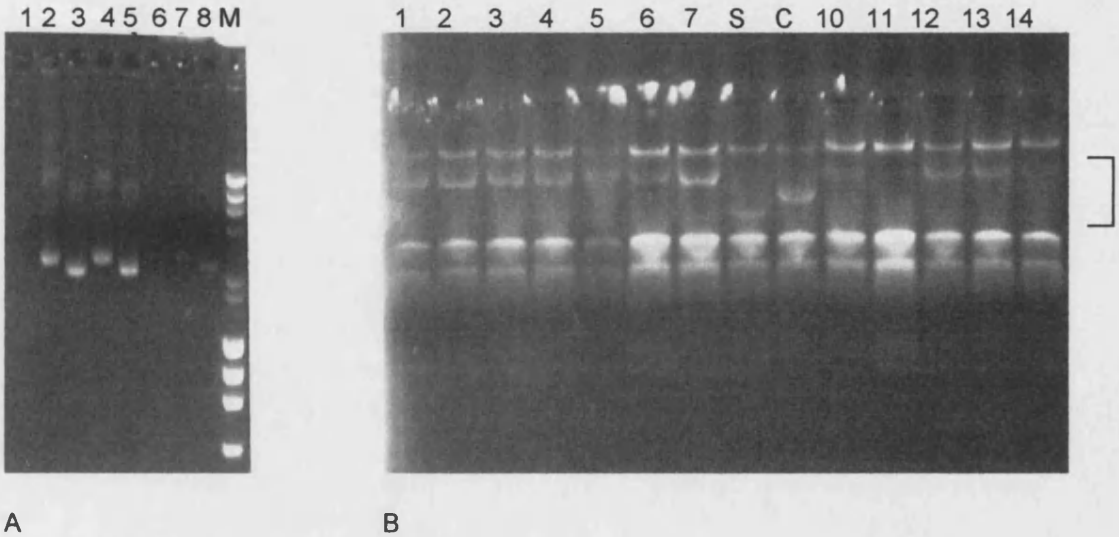


Figure 6.3 Electrophoresis at $5Vcm^{-1}$ in 1% (w/v) agarose for A: Samples from minipreps of putative recombinants for pTAg-A24 (lanes 1–8), λ *Hin*dIII/ϕX-174 *Hae*III molecular weight markers (M). B: Single colony gel analysis results for pTAg-A30 putative recombinants. S: self-ligated plasmid; C: plasmid with control insert. Band variation was observed in the area indicated by the bracket.

6.3.1.2 *Identification of the clones by DNA sequencing*

Approximately 110bp sequencing runs were obtained for the clones of A24, while 190bp and 120bp were obtained for the A30 forward and reverse reactions, respectively. Comparison of these preliminary sequencing results for the two sequences with the databases, using the BLAST and FASTA programmes, identified them as ribosomal DNA sequences for the large (26S) subunit. Further data were obtained using automated sequencing reactions, with 500–650bp of sequence typically obtained. Since A24 was small, the whole of the insert was sequenced in both directions using these primers. However, A30 was over 1kb and internal primers were therefore designed to confirm sequence that was ambiguous or not covered by the results so far. Figure 6.4 shows the position of these primers, their directions and sizes.

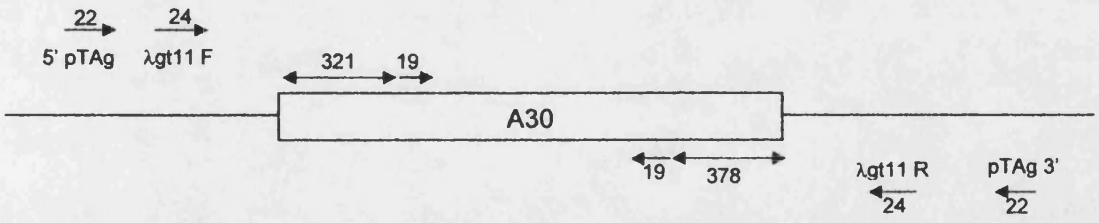


Figure 6.4 Schematic diagram of the positions of the primers used to sequence A30, with respect to the cloning site within the pTAg vector (not to scale), and their length given in bp. Positions of the internal primers designed against A30 are shown relative to the ends of the A30 sequence. Primers located within the two vectors pTAg and λgt11 are applicable for the sequencing of pTAg-A24 also.

The resulting sequences enabled further BLAST searches to be done, and good similarity to the large subunit DNA sequences from a wide range of plant species was obtained.

A pairwise alignment of A24 and A30 showed that the part of the large subunit sequence represented by clone A24 was located within that encoded by A30. Therefore A30 was used for the subsequent phylogenetic analyses. Figure 6.5 shows the full nucleotide sequence of A30, aligned with the 26S ribosomal DNA sequence from *Citrus limon*, which gave the highest percentage similarity over the full sequence submitted from the BLAST search. Other plant 26S rDNA sequences are approximately 3.4kb in full (Hamby and Zimmer, 1992) so it appears that A30 represents around a third of the full cassava 26S sequence.

Figure 6.5 Partial nucleotide sequence (1169bp) of the cassava 26S ribosomal DNA sequence, clone A30, aligned with the corresponding region from *Citrus limon*. Of 1169bp, 1134 (97%) were identical. Nucleotide positions which vary between the two sequences are highlighted.

```
Cassava   CTTCCCCGGGCGTCGAACAGTCGACTCAGAACTGGTACGGACAAGGGGAATCCGACTGTTTAATTAAAACAAAGCATTGCGATGGTCCCTGCGGATGCTC   100
C.limon   CTTCCCCGGGCGTCGAACAGTCGACTCAGAACTGGTACGGACAAGGGGAATCCGACTGTTTAATTAAAACAAAGCATTGCGATGGTCCCTGCGGATGCTC

Cassava   ACGCAATGTGATTTCTGCCCAGTGCTCTGAATGTCAAAGTGAAGAAATTCAACCAAGCGCGGGTAAACGGCGGGAGTAACTATGACTCTCTTAAGGTAGC   200
C.limon   ACGCAATGTGATTTCTGCCCAGTGCTCTGAATGTCAAAGTGAAGAAATTCAACCAAGCGCGGGTAAACGGCGGGAGTAACTATGACTCTCTTAAGGTAGC

Cassava   CAAATGCCTCGTCATCTAATTAGTGACGCGCATGAATGGATTAACGAGATTCCCACTGTCCCTGTCTACTATCCAGCGAAACCACAGCCAAGGGAACGGG   300
C.limon   CAAATGCCTCGTCATCTAATTAGTGACGCGCATGAATGGATTAACGAGATTCCCACTGTCCCTGTCTACTATCCAGCGAAACCACAGCCAAGGGAACGGG

Cassava   CTTGGCRGAATCAGCGGGGAAAGAAGACCCTGTTGAGCTTGACTCTAGTCCGACTTTGTGAAATGACTTGAGAGGTGTAGRATAAGTGGGAGCCGGAAAC   400
C.limon   CTTGGCAGAATCAGCGGGGAAAGAAGACCCTGTTGAGCTTGACTCTAGTCCGACTTTGTGAAATGACTTGAGAGGTGTAGTATAAGTGGGAGCCGGAAAC

Cassava   GGCGACYGTGAAATACCACTACTTTTAACGTTATTTTACTTATTCCGTGAATCGGAGGCGGGGCTTYGCCCCTCTTTTTGGACCCAMG-GCCGKCACGGC   500
C.limon   GGCGAAAGTGAAATACCACTACTTTTAACGTTATTTTACTTATTCCGTGAATCGGAGGCGGGGCACTGCCCCTCTTTTTGGACCCAGCCGCCTCACGGG

Cassava   GGCCGATCCGGGCGGAAGACATTGTCAGGTGGGGAGTTTGGCTGGGGCGGCACATCTGTTAAAAGATAACGCAGGTGTCCTAAGATGAGCTCAACGAGAA   600
C.limon   GGCCGATCCGGGCGGAAGACATTGTCAGGTGGGGAGTTTGGCTGGGGCGGCACATCTGTTAAAAGATAACGCAGGTGTCCTAAGATGAGCTCAACGAGAA

Cassava   CAGAAATCTCGTGTGGAACAAAAGGGTAAAAGCTCGTTTGATTCTGATTTTCAGTACGAATACGAACCGTGAAAGCGTGGCCTATCGATCCTTTAGACCT   700
C.limon   CAGAAATCTCGTGTGGAACAAAAGGGTAAAAGCTCGTTTGATTCTGATTTTCAGTACGAATACGAACCGTGAAAGCGTGGCCTATCGATCCTTTAGACCT

Cassava   TCGGAATTTGAAGCTAGAGGTGTCAGAAAAGTTACCACAGGGATAACTGGCTTGTGGCAGCCAAGCGTTCATAGCGACGTTGCTTTTTGATCCTTCGATG   800
C.limon   TCGGAATTTGAAGCTAGAGGTGTCAGAAAAGTTACCACAGGGATAACTGGCTTGTGGCAGCCAAGCGTTCATAGCGACGTTGCTTTTTGATCCTTCGATG

Cassava   TCGGCTCTTCCTATCATTGTGAAGCAGAATTCACCAAGTGTTGGATTGTTCACCCACCAATAGGGAACGTGAGCTGGGATTAGACCGTCGTGAGACAGGT   900
C.limon   TCGGCTCTTCCTATCATTGTGAAGCAGAATTCACCAAGTGTTGGATTGTTCACCCACCAATAGGGAACGTGAGCTGGGTTTAGACCGTCGTGAGACAGGT

Cassava   TAGTTTTACCCTACTGATGACMGCGTCGYCGATWGTAATTCAACCTAGTACGAGAGGAACCGTTGATTCGCACAATTGGTCATCGCGCTTGGTTGAAAAGC  1000
C.limon   TAGTTTTACCCTACTGATGACAGCGTCGTAATAGTAATTCAACCTAGTACGAGAGGAACCGTTGATTCGCACAATTGGTCATCGCGCTTGGTTGAAAAGC

Cassava   CAGTGGCGCGAAGCTACCGTGCGCTGGATTATGACTGAACGCCTCTAAGTCAGAATCCGGGCYAGAAGCGACGCYTGTGTCCGCCGCCCGTTTGCCGACC  1100
C.limon   CAGTGGCGCGAAGCTACCGTGCGCTGGATTATGACTGAACGCCTCTAAGTCAGAATCCGGGCTAGAAGCGACGCTTGCGCCCGCCGCCCGTTTGCCGACC

Cassava   CRCAGTAGGGGCCGTCCGGCCCCCAAAGGCACGTGCCGTTGGCCAAGCCCWCGCGGCGGACGAGCCGCG   1169
C.limon   CGCAGTAGGGACCTCCGGTCCCCAGAGGCACGTGTCGTAGGCCAAGCCCTCGCGGCGGAAGGGCCGCG
```

6.3.2 Identification of a clone for the cassava 18S ribosomal subunit

6.3.2.1 Sub-cloning and sequencing

E30 was identified by sequencing the PCR product directly. E30 was then cloned and primers within the vector used to obtain further data. Approximately 600bp of reliable sequence data were obtained from each sequencing reaction, in both directions, which gave the full extent of E30 i.e. 686bp.

The usual method of identification by comparison using BLAST and FASTA analyses showed that E30 was similar to the small (18S) ribosomal subunit from a number of plant species. Figure 6.6 gives the alignment with the sequence from *Euphorbia pulcherrima*, which was 99% identical with the cassava sequence over the E30 sequence. The full length cassava sequence would be expected to be around 1.8kb, as for other plant 18S sequences (Hamby and Zimmer, 1992).

Figure 6.6 Partial nucleotide sequence (686bp) of the cassava 18S ribosomal DNA sequence, clone E30, aligned with the corresponding region from *Euphorbia pulcherrima*. Of 686bp, 679 (99%) were identical. Nucleotide positions which vary between the two sequences are highlighted.

```
Cassava          CAAATTTCTGCCCTATCAACTTTCGATGGTAGGATAGAGGCCTACCATGGTGGTGACGGGTGACGGAGAATTAGGGTTCGATTCCGGAGA      90
E.pulcherrima    CAAATTTCTGCCCTATCAACTTTCGATGGTAGGATAGAGGCCTACCATGGTGGTGACGGGTGACGGAGAATTAGGGTTCGATTCCGGAGA

Cassava          GGGAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCAATCCTGACACGGGGAGGTAGTGACAATAAATA     180
E.pulcherrima    GGGAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCAATCCTGACACGGGGAGGTAGTGACAATAAATA

Cassava          ACAATACCGGGCTCATTCGAGTCTGGTAATTGGAATGAGTACAATCTAAATCCCTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAG     270
E.pulcherrima    ACAATACCGGGCTCATTGAGTCTGGTAATTGGAATGAGTACAATCTAAATCCCTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAG

Cassava          CAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTTAAGTTGTTGCAGTTAAAAAGCTCGTAGTTGGACCTTGGGTTGGGTCGACCGGTC     360
E.pulcherrima    CAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTTAAGTTGTTGCAGTTAAAAAGCTCGTAGTTGGACCTTGGGTTGGGTCGACCGGTC

Cassava          CGCCTCGCGGTGTGCACCTGTCGGCTCGTCCCTTCTGCCGGCGATGCGCTCCTGGCCTTAATTGGCCGGGTCGTGCCTCCGGCGCTGTTA     450
E.pulcherrima    CGCCTTTTGGTGTGCACCTGTCGGCTCGTCCCTTCTGCCGGCGATGCGCTCCTGGCCTTAATTGGCCGGGTCGTGCCTCCGGCGCTGTTA

Cassava          CTTTGAAGAAATTAGAGTGCTCAAAGCAAGCCTACGCTCTGTATACATTAGCATGGGATAACATCATAGGATTTCGGTCCTATTCTGTTG     540
E.pulcherrima    CTTTGAAGAAATTAGAGTGCTCAAAGCAAGCCTACGCTCTGTATACATTAGCATGGGATAACATCATAGGATTTCGGTCCTATTCTGTTG

Cassava          GCCTTCGGGATCGGAGTAATGATTAACAGGGACAGTCGGGGGCATTCGTATTTCATAGTCAGAGGTGAAATTCTTGGATTTATGAAAGAC     630
E.pulcherrima    GCCTTCGGGATCGGAGTAATGATTAACAGGGACAGTCGGGGGCATTCGTATTTCATAGTCAGAGGTGAAATTCTTGGATTTATGAAAGAC

Cassava          GAACAACTGCGAAAGCATTTGCCAAGGATGTTTTCATTAATCAAGAACGAAAGTTG                                     686
E.pulcherrima    GAACAACTGCGAAAGCATTTGCCAAGGATGTTTTCATTAATCAAGAACGAAAGTTG
```

6.3.3 Comparison of transcription units from different cultivars

Restriction fragment length polymorphism (RFLP) analysis was done for several cultivars, the gel processed as described for Southern analysis and probed under stringent conditions with radiolabelled A30. Genomic DNA samples from several cultivars were prepared and 5μg digested singly with *Hin*dIII, *Eco*RI and *Eco*RV, these three enzymes chosen since they reportedly cut only once per transcription unit (Fregene *et al.*, 1994). Figure 6.7 shows the autoradiograph of the Southern blot, using 5.5h exposure at –70°C. Two further exposures of 2.5h and 24h were used to help resolve bands accurately (not shown).

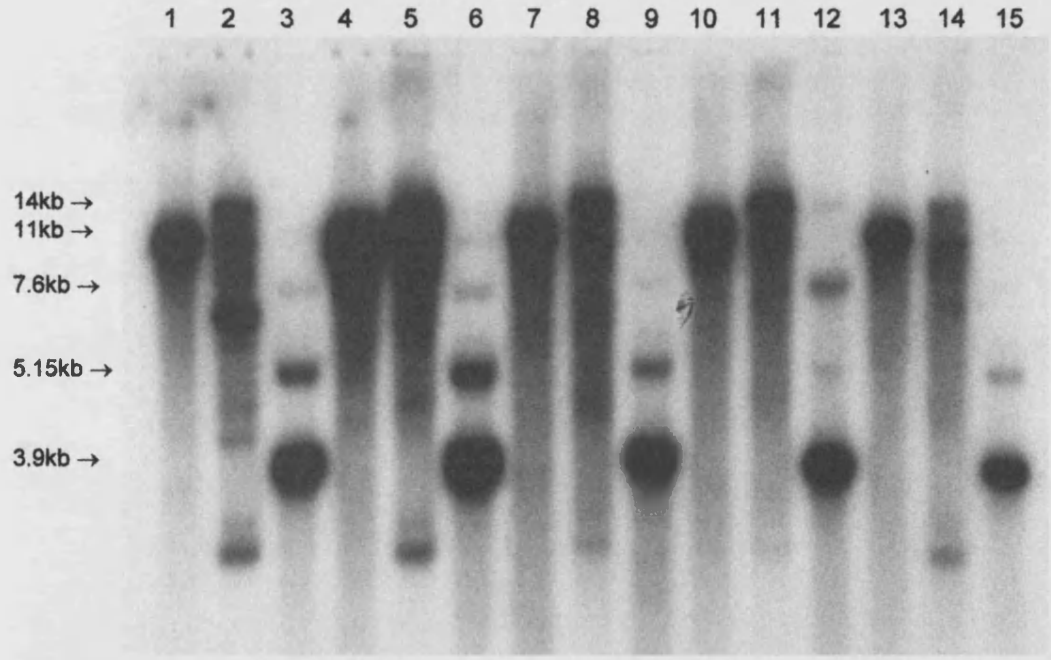


Figure 6.7 Autoradiograph developed at –70°C for 5.5 h. Lanes 1–3 contain digests for CMC40, 4–6 MCol 1684, 7–9 NGA2, 10–12 MVen 77 and 13–15 MBra 337. Digests for each cultivar were loaded in the order *Hin*dIII, *Eco*RV and *Eco*RI. The gel was run at 1.5Vcm^{-1} overnight at room temperature, blotted as described previously and probed with 30ng of a purified *Bam*HI fragment of A30 (single *Bam*HI sites are located in each adaptor of the cloned A30 sequence).

All cultivars gave the same result for the *Hin*dIII digest, namely the expected single band, at approximately 11kb. However, several bands were obtained for the *Eco*RI and *Eco*RV digests. Differences can be seen both in the profile of CMC 40 digested with *Eco*RV, and MVen 77 digested with *Eco*RI, when compared with the other cultivars. CMC 40 has 5 bands in common with the other cultivars, bands 1, 2, 3, 6 and 8 at approximately 14kb, 9.4kb, 6.7kb, 4.7kb and 3.2kb respectively. Additional bands are apparent at 7.8kb, 5.8kb and 4.2kb in the CMC 40 track.

Four Bands appear in the MVen 77 lane at approximately 7.6kb, 5.15kb and 3.9kb, as for the other cultivars, but only the 3.9kb band is of a similar relative intensity. The band at 7.6kb is more intense and that at 5.2kb less intense than the corresponding bands for the other cultivars. Additionally, the highest molecular weight band for MVen 77 is approximately 14kb compared with bands of 11kb for the remaining cultivars.

Fregene and coworkers (1994) reported a second band for MBra 337 digested with *Hin*dIII, given a long exposure time, but no second band was apparent here, even for the overexposed 24h autoradiograph.

6.3.4 Phylogenetic analysis

The two clones A30 and E30 were used for phylogenetic analysis. DNA sequences for both the large and the small ribosomal subunits from a number of plants were obtained from the respective databases on the internet at the University of Antwerp (De Rijk *et al.*, 1999; Van de Peer *et al.*, 1999).

Relatively few plant sequences for the 26S sequence, or large ribosomal subunit DNA, were available for comparison with the cassava sequence. Those used in the analysis are listed in Table 6.1, together with their common names.

Table 6.1 List of those flowering plant partial sequences for the large ribosomal subunit used for phylogenetic comparison with the corresponding ribosomal sequence isolated from cassava (*Manihot esculenta*). *Chlorella elipsoidea* was used as the outgroup

Plant ribosomal sequence	Accession number	Common name
Monocotyledons		
Oryza sativa	M11585	rice
Dicotyledons		
Arabidopsis thaliana	X52320	thale cress
Brassica napus	D10840	oilseed rape
Citrus limon	X05910	lime
Fragaria ananassa	X58118	strawberry
Lycopersicon esculentum	X13557	tomato
Medicago sativa	Z11498	alfalfa
Sinapis alba	X57137	white mustard
Outgroup		
Chlorella ellipsoidea	D17810	

Many more sequences were available for the small, or 18S, subunit DNA sequence, and those used are listed in Table 6.2. In addition to a few representatives from the spore-bearing plants and three from the gymnosperms, these were chosen to provide flowering plant sequences representing a wide range of families. The four sequences found from other members of the Euphorbiaceae (*Adriana glabrata*, *Drypetes roxburghii*, *Euphorbia pulcherrima* and *Ricinis communis*) were included and where possible, plants whose large subunit gene sequences were used were included in the small subunit analysis also. In both cases, the sequence for the appropriate subunit from blue-green algae (*Chlorella spp.*) was used as an outgroup, since divergence of algae from higher plants occurred before the speciation of the Euphorbiaceae.

Table 6.2 Details of plant small ribosomal subunit DNA sequences used in comparison with the cassava partial sequence. Common names of individuals, or those of the genus to which the individual species belong, are indicated where possible (Benson, 1979; Mabberley, 1997)

Plant ribosomal sequence	Accession number	Common individual (or genus*) name
Bryophyta		
Marchantia polymorpha	X75521	liverwort
Trachaeophyta		
Horsetails		
Equisetum hyemale	U18500	Dutch rush
Psilophytes		
Psilotum nudum	X81963	
Clubmosses		
Sphagnum cuspidatum	X80213	liver moss
Gymnosperms		
Ginkgo biloba	D16448	maidenhair tree
Gnetum leyboldii	L24045	
Pinus elliottii	D38245	pine*
Angiosperms		
Adriana glabrata	L37581	bitterbush*
Alnus glutinosa	X54984	alder*
Aristolochia tomentosa	L24083	birthwort*
Australobaileya scandens	U38219	
Balanophora fungosa	L24044	
Buxus sempervirens	X16599	box*
Citrus aurantium	U38312	sour orange
Cornus racemosa	X16602	dogwood*
Cuscuta gronovii	L24747	dodder*
Drypetes roxburghii	U42534	
Eleagnus umbellata	L24090	oleaster*
Euonymus alatus	X16600	burning bush*
Euphorbia pulcherrima	L37582	poinsettia
Fragaria ananassa	X15590	strawberry
Glycine max	X02623	soyabean
Gossypium hirsutum	L24145	cotton
Hedera helix	X16604	ivy*
Linaria vulgaris	U38315	common toadflax
Linum perenne	L24401	flax*
Morus alba	L24398	white mulberry
Nymphaea tuberosa	L24404	water lily*
Oryza sativa	X00755	rice
Paeonia lactiflora	L24410	Chinese peony
Pittosporum japonicum	L28142	
Ranunculus sardous	L24092	Sardinian buttercup
Ribes aureum	L28143	golden/buffalo currant
Ricinis communis	L37583	castor bean
Santalum album	L24416	white sandalwood
Sinapis alba	X17062	white mustard
Solanum tuberosum	X67238	potato
Spinacia oleracea	L24420	spinach
Tropaeolum majus	L28750	nasturtium*
Zea mays	K02202	maize

As most of the sequences for both subunits were for the respective genes and comparison was being made with the cassava cDNA sequences, it was necessary to trim the sequences to approximately the same size. CLUSTALW (Thompson *et al.*, 1994) was therefore used to align the two groups of sequences to show the areas of overlap.

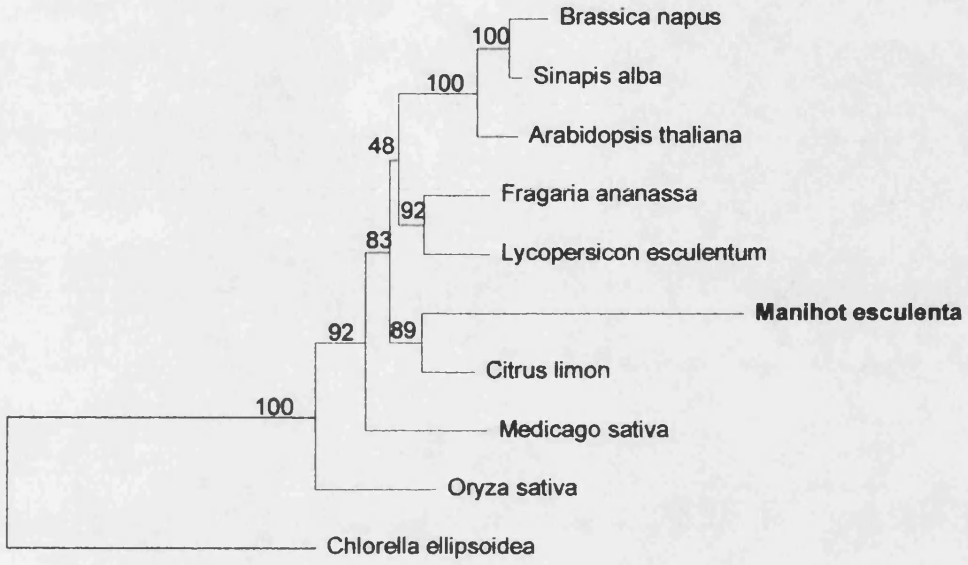
The data were analysed using two different methods. Neighbour-joining, using the TREECON programme (version 1.3b; Van de Peer *et al.*, 1993), was used as an example of a distance method, as well as DNAPARS (part of the phylogenetic inferencing, or PHYLIP, package, version 3.5c; Felstenstein, 1993), which uses maximum parsimony. This latter method uses a CLUSTALW alignment file as the input, but the option to randomize the input order of the interleaved sequences was taken. Both methods were bootstrapped, which involves resampling from the data set and assessing the frequency with which particular groups, or clades, appear in the trees generated from this resampled data.

6.3.4.1 26S ribosomal sequence

Figure 6.8 shows the output for the large subunit sequences analysed by the TREECON neighbour-joining method. The tree produced by using the Kimura correction is shown, as this weights transitions and transversions differently and it was assumed that for the highly conserved, slowly evolving rDNA sequences this would be important. A tree generated with the Jukes and Cantor correction, which compensates for multiple substitutions at the same site, gave an identical pattern to the Kimura tree, with similar bootstrap values (not shown).

Figure 6.9 gives the consensus tree obtained by running the CONSENSE programme on the outfile from the DNAPARS maximum parsimony (advanced) programme.

Figure 6.8 Tree generated for A30, the cassava sequence for the 26S ribosomal subunit, when compared with bp2085–3254 of the corresponding sequences from eight other flowering plant and one algal sequence, *C. ellipsoidea*, which was used as the outgroup to root the tree. Analysis was carried out using the TREECON programme, with the Kimura correction and 100 bootstraps, values for which are shown at each branch point. The position of cassava within the tree is highlighted.



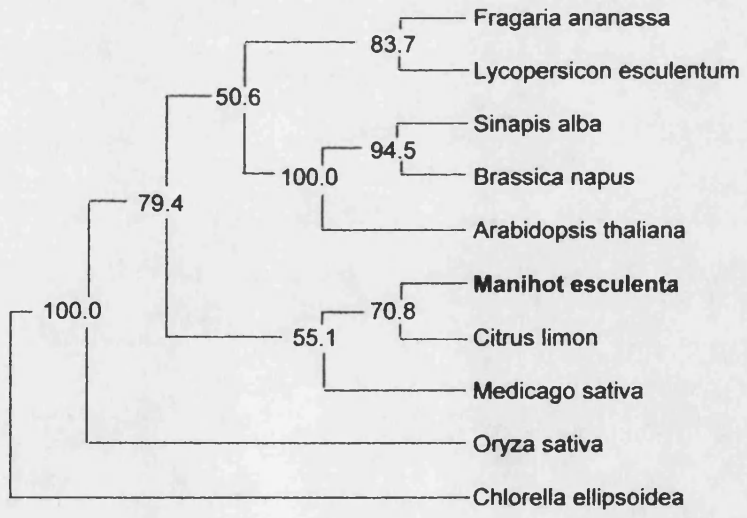


Figure 6.9 Tree generated for A30, the cassava sequence for the 26S ribosomal subunit, when compared with bp2085–3254 of the corresponding sequences from eight other flowering plant and one algal sequence, *C. ellipsoidea*, which was used as the outgroup to root the tree. Analysis was carried out using the outfile from DNAPARS in the CONSENSE programme., with 100 bootstraps and the input order randomized. Numbers at the forks indicate the number of times the group consisting of the species which are to the right of that fork occurred among the trees, out of 100 trees.

6.3.4.2 18S ribosomal sequence

Figures 6.10 and 6.11 show the output for the small subunit sequences analysed by TREECON and DNAPARS, as for the 26S sequences. Figure 6.10 again shows the neighbour-joining results from the Kimura correction. As with the large subunit sequence, a similar (though not identical) tree was also produced with the Jukes and Cantor correction (not shown). Where there was good confidence with the groupings, these were obtained in both trees, thus the bottom parts of both trees from *Chlorella* to *Balanophora* were identical. *Zea* and *Oryza*, and *Ricinis* and *Adriana* both grouped together in similar position. Differences were seen in the positioning of some of the upper parts of the tree, but as the bootstrap values here are low, it was decided only to include the Kimura tree.

Figure 6.11 gives the consensus tree obtained by running the CONSENSE programme on the outfile from the DNAPARS maximum parsimony (advanced) programme, with a CLUSTALW alignment as the original input file. The input order was randomized and 100 bootstraps done, with numbers at the forks indicating the number of times the group consisting of the species to the right of the fork occurred among the trees, out of 100 trees.

Lineages for the sequences used are shown in Figures 6.12–6.14. These were constructed by comparing the information accompanying the sequences in GenBank form, with additional information obtained from Benson (1979). Figure 6.14 continues the lineage of the Eudicotyledons, established in Figure 6.13.

Figure 6.10 E30, the partial sequence for the cassava 18S ribosomal subunit, compared with the corresponding region from 40 other flowering plants and the algae *Chlorella zofingiensis* (chosen as the outgoup) using the TREECON programme (Kimura correction). One hundred bootstraps were carried out during the generation of the tree, numbers for which are shown on the branch points.

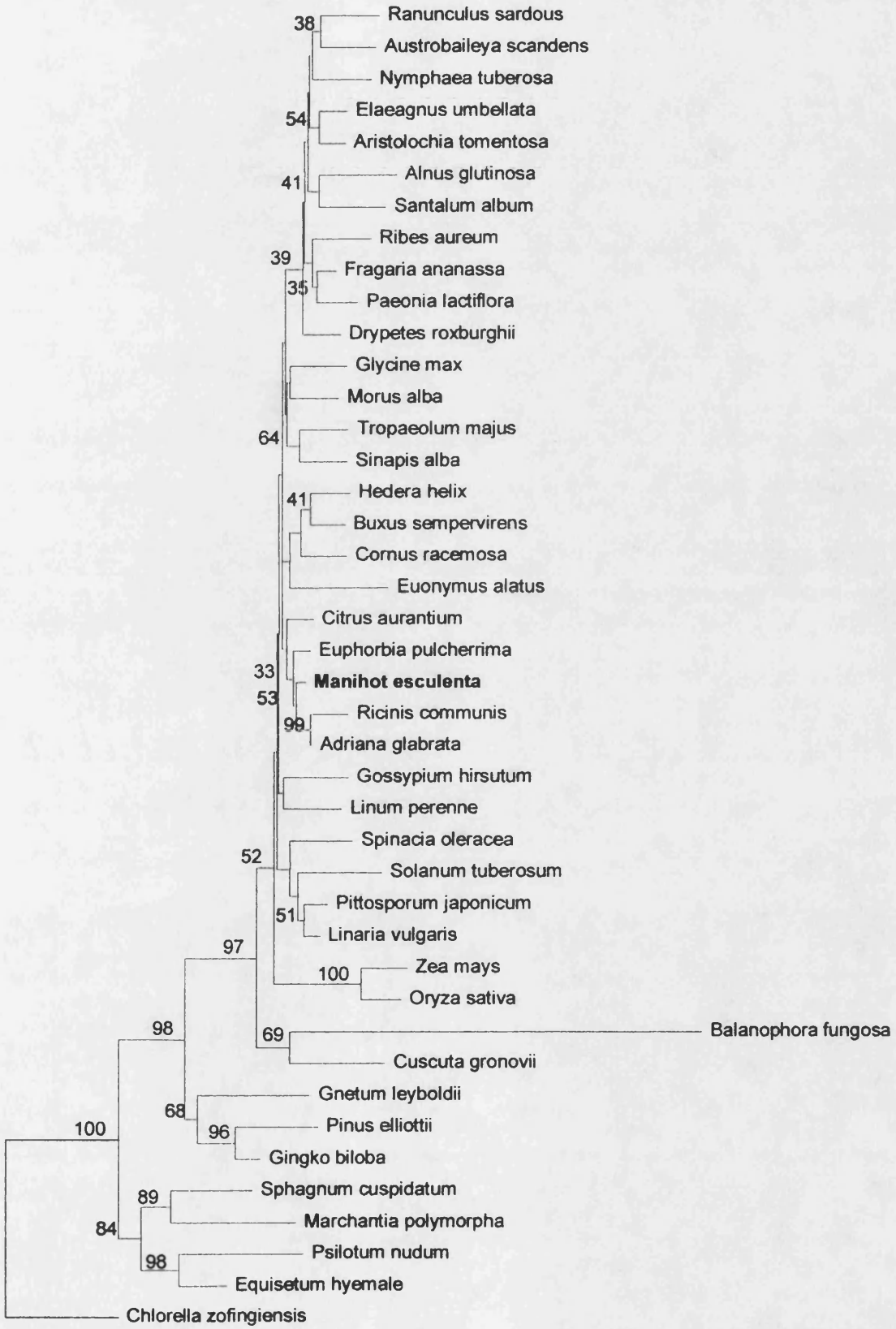
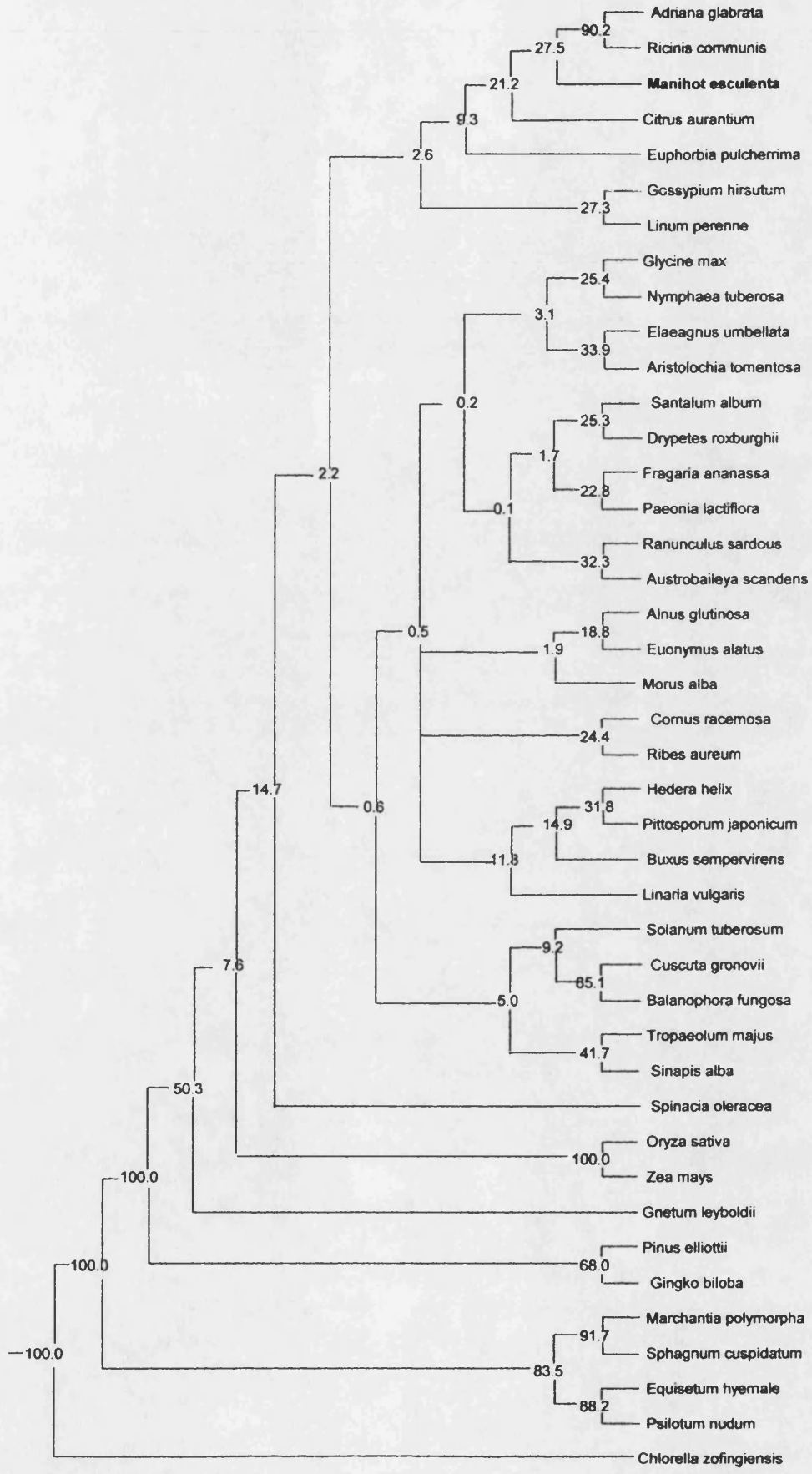


Figure 6.11 E30, the partial sequence for the cassava 18S ribosomal subunit, compared with the corresponding region from 40 other flowering plants and the algae *Chlorella zofingiensis* (chosen as an outgoup) using maximum parsimony. One hundred bootstraps were carried out, numbers for which are shown on the branch points.



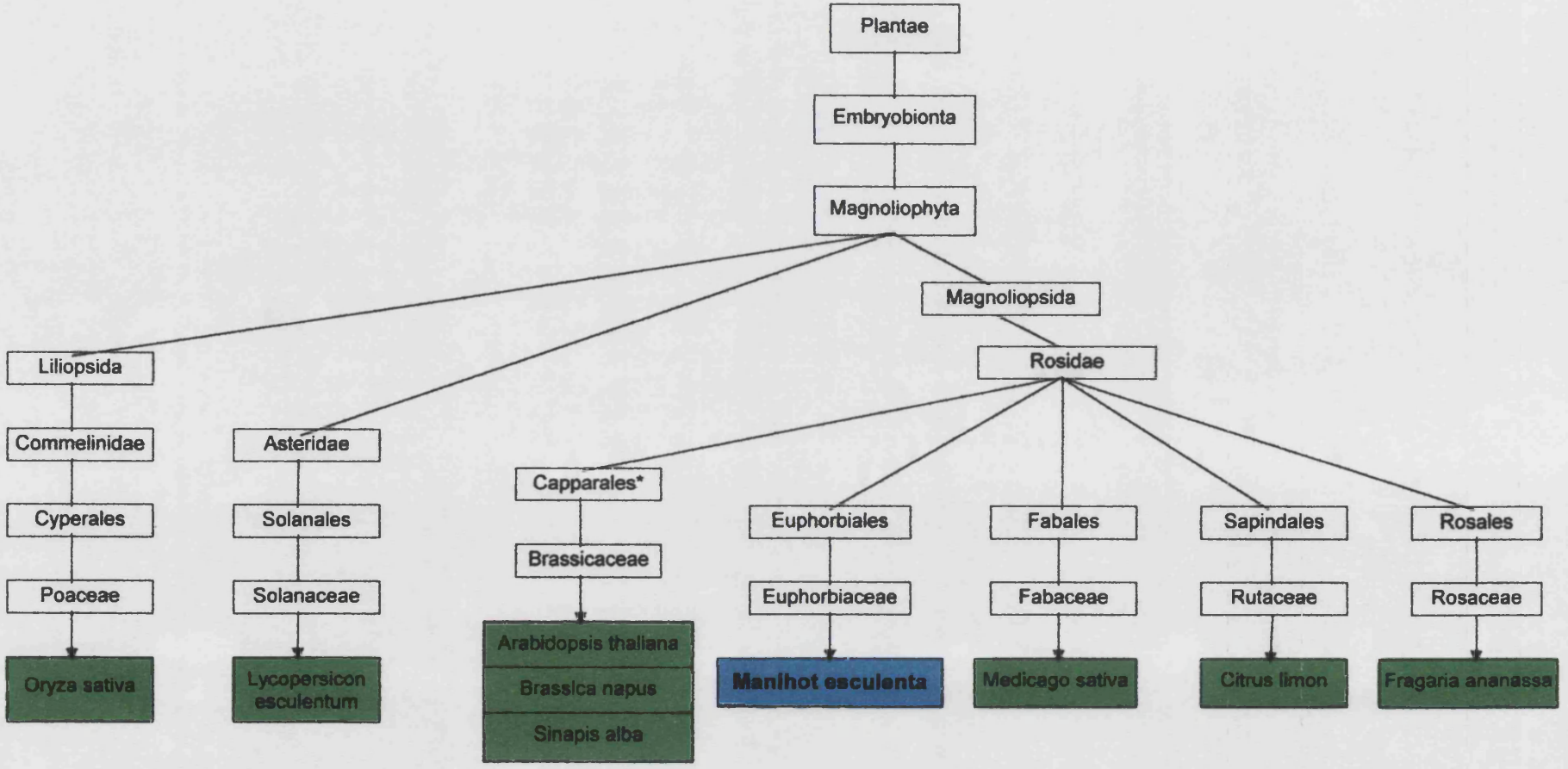


Figure 6.12 Flow diagram showing the lineages of the nine flowering plants used in phylogenetic analysis with the 26S ribosomal subunit partial sequence from cassava. Lineages are based on the information accompanying each of the database accessions. *Formerly Dilleniidae.

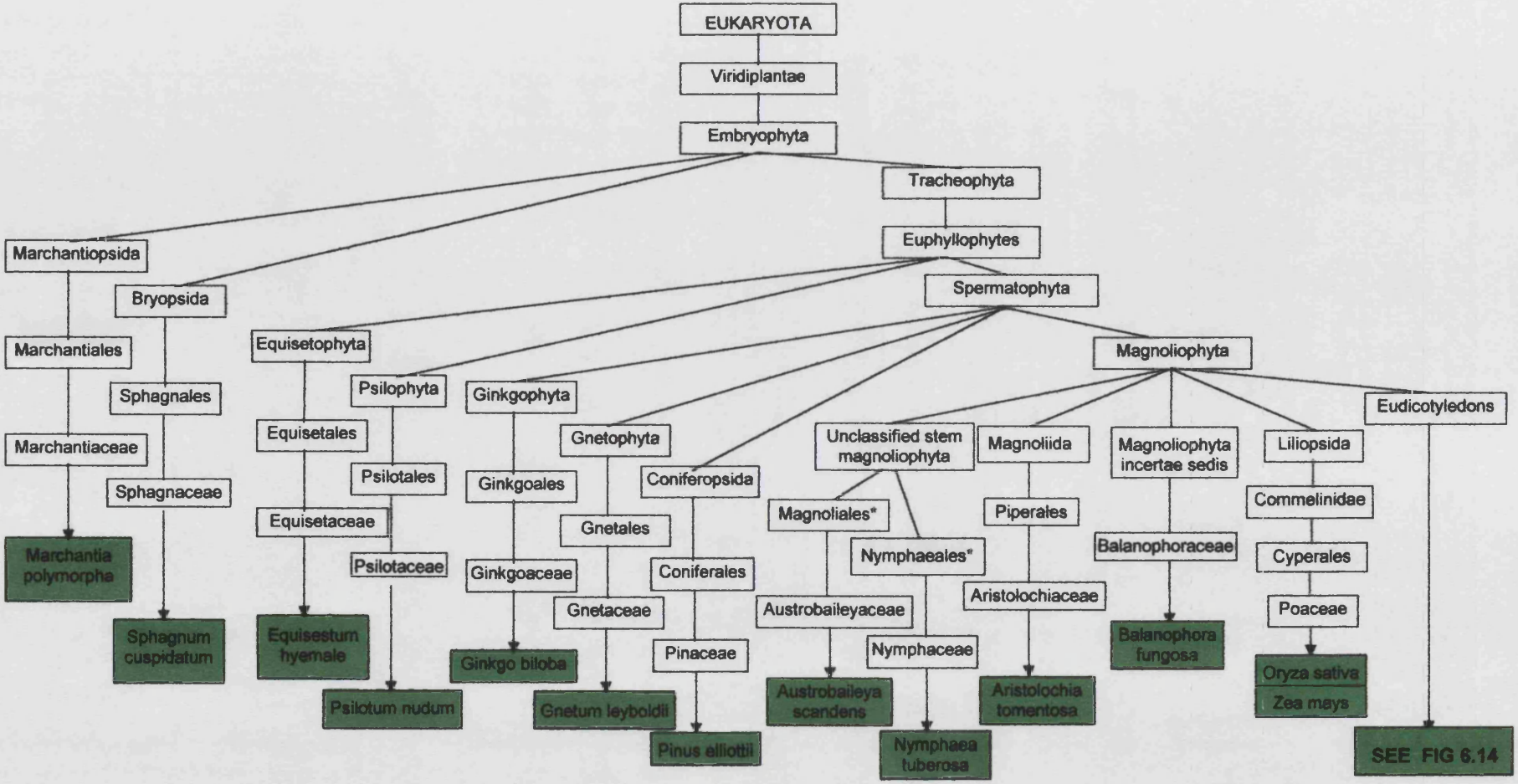


Figure 6.13 Classical lineages constructed using the information accompanying the database accessions of the individual species (highlighted) whose 18S ribosomal subunit DNA sequences were used in the phylogenetic analysis with the partial cassava sequence. *Formerly Magnoliidae.

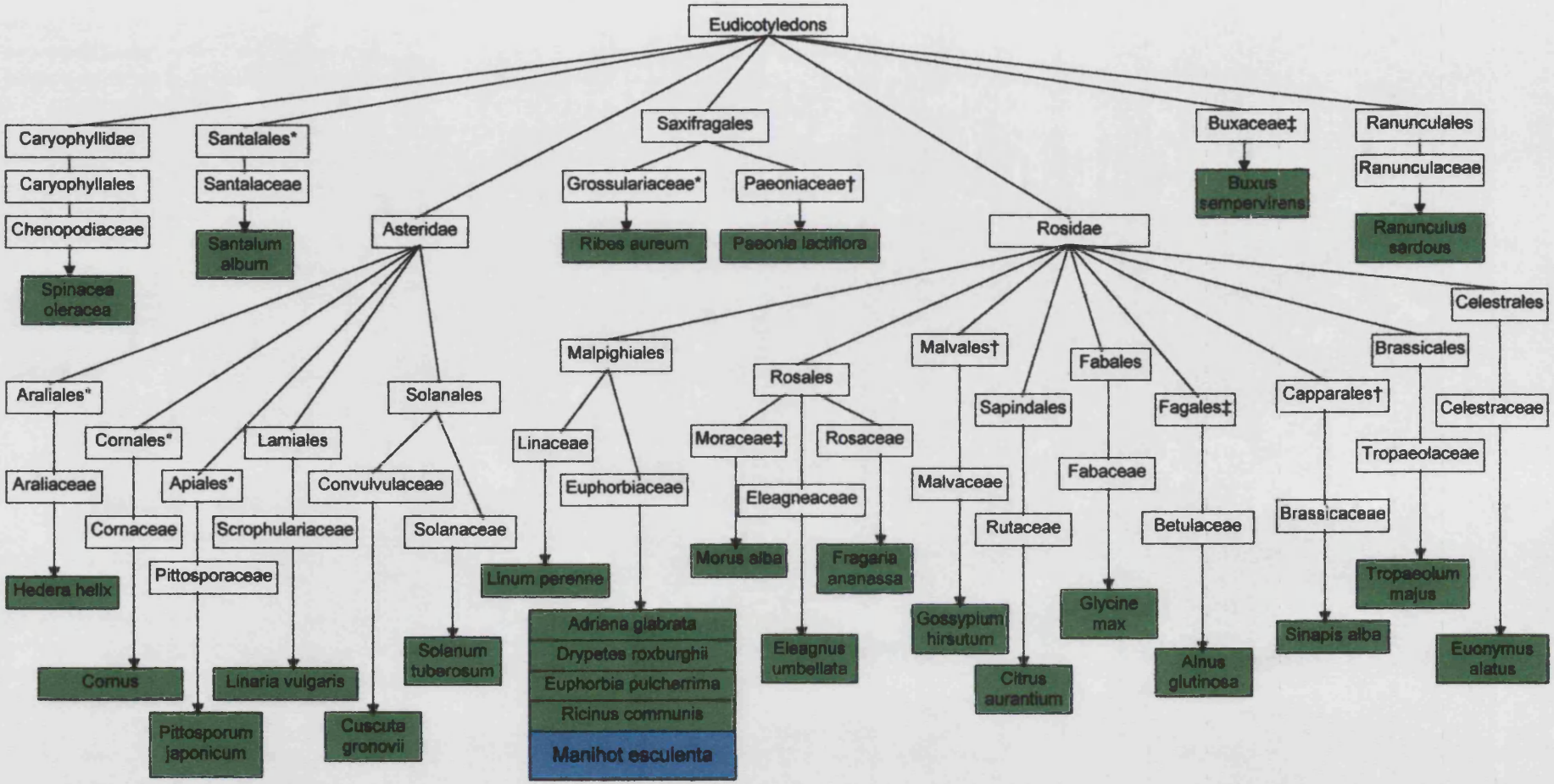


Figure 6.14 Continuation from Figure 6.13 of the classification lineages for those eudicotyledon species whose 18S ribosomal subunit DNA sequences were used for phylogenetic analysis with the partial cassava sequence. *Formerly Rosidae; †Formerly Dilleniidae; ‡Formerly Hamamelidae.

The study by Chase *et al.* (1993) on the phylogenetics of the seed plants using the plastid gene *rbc*L was used for comparison with the results from the cassava analyses. A number of differences were noticed between the lineages as cited by Chase *et al.* (1993) and the lineages stated in the GenBank information. It has been assumed that the 1993 study has been used to reclassify a number of families, information which has subsequently been incorporated within the GenBank accession details. These differences have been indicated in Figures 6.12–6.14. For example, the lineage accompanying the database accession for *Cornus* lists it as a member of the Asteridae. Chase and co-workers refer to the Cornaceae as being considered to belong to the Rosidae, but their 1993 study placed them in Asterid clade IV. Similarly, the database classifies the Brassicaceae as Rosidae, whereas formerly they have been classified as Dilleniidae. Chase *et al.* (1993) placed the Brassicaceae within Rosid clade II.

6.4 DISCUSSION

6.4.1 Variation between cultivars in rDNA repeats

The results from the Southern analysis of the digested gDNA were as expected for the *Hin*dIII enzyme. This gave a single band of 11kb for all five cultivars indicating that it cut once per repeat unit. This was the result obtained by Fregene *et al.* (1994), and is consistent with the observation of Hamby and Zimmer (1992) that plant transcription rDNA repeat units are between 9kb and 11kb in length. The second band reported by Fregene and coworkers in MBra 337 after extended exposure time, indicating a variation in IGS length, was not seen here.

The patterns for *Eco*RV and *Eco*RI were unexpected. For *Eco*RV, the highest molecular weight band for all cultivars is equivalent to 14kb. This would appear to indicate multiple sites within the transcription unit, together with partial digestion. As

the subsequent sites are cut the smaller bands are released. This would explain the many bands obtained for all cultivars and the eventual release of the band at 3150bp. CMC 40 gave a different pattern with *Eco*RV, although it does share several bands, which may indicate that at least one site is different.

The profiles obtained with *Eco*RI were also not as anticipated assuming a single restriction site, indicating more sites for this enzyme also. The difference in intensities between the bands at 7.6kb and 5.2kb, when MVen 77 and the other cultivars are compared, seems to indicate partial digestion and the presence of two further *Eco*RI sites. The whole transcription unit appears to be cut at site one to produce the bands at approximately 3.9kb and 7.6kb, with possible subsequent digestion of the latter to release the 5.2kb band. The appearance of a faint band at 11kb, equivalent to the full transcription repeat unit length, is likely to result from the presence of an *Eco*RI site in A30, the probe used. Since the *Eco*RI site is positioned at 827–832bp within A30, enough of the probe is available to hybridize to both fragments, effectively religating them. Since the 3.9kb band is very intense in all cases, it is unlikely that partial digestion explains the 11kb band. The highest molecular weight band in MVen 77 was different for this digestion, being 14kb rather than 11kb. This is consistent with the alternative 'full length' transcription unit variant obtained for CMC 40 with *Eco*RV, possibly due to IGS length variation. Since these genes are thought to exhibit concerted evolution by means of unequal cross-over and gene conversion, a mutation causing an extra restriction site would be expected to be rapidly spread throughout the repeats, resulting in bands being visible rather than a smear. The *Eco*RV digest for CMC 40 may reflect a site being present in some units and not others – possibly a more recent mutation, or a variant within the less conserved IGS.

6.4.2 Analysis of the cassava 26S ribosomal sequence

Results from the two methods of analysis were similar for the 26S rDNA sequences. Generally the bootstrap values were high, indicating good confidence for the groupings shown. In both cases the sequence for rice was the first to branch off, with reasonable confidence. This is the only monocot within the group used, so this result is as expected. Neighbour-joining then predicts divergence of the sequence for alfalfa, with the remaining sequences forming two groups 83% of the time. The first contains cassava and lime, and the other rape, mustard, *Arabidopsis*, strawberry and tomato. The first three sequences are all members of the Brassicaceae and group together 100% of the time, rape and mustard being more similar to each other than to *Arabidopsis*. Strawberry and tomato are most similar, but the bootstrap value for them to form a separate group from the Brassicaceae is low. BLAST results gave the sequence for lime as most similar to that for cassava, so it is reassuring to see them placed together.

The groupings predicted by parsimony are very similar, with the position of alfalfa within the group containing lime and cassava the only difference. There is a little more confidence for the separation of strawberry and tomato from the Brassicaceae; just over 50% compared with 48% for neighbour-joining. The only surprising result is the positioning of strawberry and tomato together with good confidence. Based on previous classification lineages (Figure 6.12), strawberry would be expected to group with alfalfa, lime and cassava, the other Rosidae, rather than tomato, a member of the Asteridae. To investigate whether this was a consequence of analysing only part of the 26S sequence, further analysis was done using the Phylip package (as previously described) on the available full length sequences. The results are shown in Appendix 4. This time, tomato and lime are grouped with more confidence than tomato and strawberry indicating that full length analysis is more reliable for this group. As with the two partial sequence analyses, rape and mustard are more similar to each other than to *Arabidopsis*.

6.4.3 18S rDNA sequence results

The lower parts of the two arrangements of the small ribosomal subunit DNA sequences are very similar, as expected. With *Chlorella* as the outgroup, the next group of sequences are those for the vascular non-flowering plants. *Equisetum* and *Psilotum* are similar, as are *Sphagnum* and *Marchantia*. These are followed by the gymnosperms, with *Ginkgo* and *Pinus* being more similar to each other than to *Gnetum*. This positioning is predicted by the study by Hamby and Zimmer (1992), as is the parsimony result for the monocots *Oryza* and *Zea*. However, the paleoherbs *Aristolochia* and *Austrobaileya*, together with *Nymphaea*, would also be expected here, according to Chase *et al.* (1993). Instead they appear scattered within the rosids. For the neighbour-joining tree, *Cuscuta* and *Balanophora* appear next, before the monocots. These group together in both analyses. *Cuscuta* is a member of the Solanaceae so it is unsurprising to see *Solanum* as the next species in the parsimony tree. *Spinacea* is placed above the monocots, according to the parsimony results, but within the group with *Solanum*, *Pittosporum* and *Linaria* for neighbour-joining. Chase *et al.* (1993) found that the Caryophillids, of which *Spinacea* is a member, grouped either within the Rosids or as a sister group basal to the Asteridae and Rosidae.

Of the remaining sequences, there are few branch points predicted with good confidence in either analysis method. Both group *Ricinis* and *Adriana* together as being most similar to each other, cassava branching off from these two, and *Citrus* and *Euphorbia* being the next two sequences within the group. *Drypetes* is the only member of the Euphorbiaceae not to group closely to the other sequences, which is unexpected. Chase *et al.* (1993) place *Drypetes* and *Euphorbia* within the same Rosid I clade. However, that they are not together here is probably a reflection of the low bootstrap values, which explains the other anomalies. Again, the reliable parts of both trees agree with previous results.

6.4.4 Future rDNA studies using cassava

Future work to identify a more reliable evolutionary position for cassava would benefit from the whole sequences being used, since different parts of the 26S and 18S rDNA evolve at different rates. Within the parts analysed here for both coding regions, more conserved and less conserved areas were seen, in accordance with previous observations (Hamby and Zimmer, 1992). However, the results using truncated sequences are consistent provided there is good confidence in the partial sequence.

Another use for the rDNA sequences would be to compare the nucleotide sequences of different cultivars. The result from the RFLP experiment indicates that multiple sites exist for *Eco*RV and *Eco*RI, enzymes reported to cut only once (Fregene *et al.*, 1994). Sequencing from different cultivars may give a better explanation for this observation and would allow analysis at the species level. Additionally, the use of genes that are more variable at the sequence level would probably be more informative.

CHAPTER 7: CHARACTERIZATION OF A NOVEL CASSAVA GENE

7.1 INTRODUCTION

This chapter deals with the characterization of clone A41 which contained an insert of 395bp, which in turn gave rise to a cDNA of 270bp. As with the other clones investigated, the first experiments were to obtain some sequence data for comparison with the databases and Southern analysis to confirm its origin as cassava. Again, the hope was that it would be an interesting sequence which could be investigated further and used to obtain the corresponding gene.

7.2 MATERIALS AND METHODS

7.2.1 PCR of cytochrome c

Amplification of cassava cytochrome c from both CMC 40 and MCol 1684 leaf genomic DNA was carried out using the second PCR amplification components and cycle outlined above for the isolation of the 'E' and 'Z' clones (see *2.3.3.5.3 PCR of λgt11 cDNA library clones*). The degenerate primers used (see Fig. 7.1) were designed against nine genomic cytochrome c sequences, including that from the plant *Arabidopsis thaliana* (kind gift of Dr. A. Bailey). The product expected was 219bp plus an intron, since *A. thaliana* has an intron of approximately 1kb at position 165 of this sequence.

Figure 7.1 Details of the forward and reverse degenerate primers used to amplify the cytochrome c sequence from cassava.

Forward primer: (cyt1)

```
5' TTT AAG ACG CGG TGT AAG GAG TGT CAT AC 3'
  **C G*A **A A*A **C GCA **A **C **C **
  *** *** **T **T *** **T *** *** *** **
  *** *** **C **C *** **C *** *** *** **
```

Reverse primer: (cyt2)

```
5' GC CAT TTT GGT GCC GGG TAT ATA TTT TTT 3'
  ** *** C** A** A** A** A** G** C** C**
  ** *** T** T** T** G** *** *** *** ***
  ** *** G** C** C** *** *** *** *** ***
```

7.2.2 A41 internal amplification

The programme PRIME (GCG) was used to design two sets of primer pairs for A41 which would give a maximum size amplification product and a smaller product, internal to this. Ideally, all four primer combinations would be possible.

The primers shown above in Table 7.1 were selected, all with optimal annealing temperatures of 50°C. Primers A and D give an expected product of 233bp, while B and C amplify 146bp.

Table 7.1 Sequences of primers designed against the original PCR fragment A41, numbers denoting the position on that fragment. F: forward primers; R: reverse primers.

Primer		Sequence 5'–3'	
A (F)	35	GTC CTG GCT TAA ATT TGC	52
B (F)	51	GCA TAG AAG TTA CAT CCA AGA G	72
C (R)	177	GGA TGA ATA TGA TGA GGA AG	196
D (R)	258	CTT ATT CGA CTT GTG GTT CT	267

7.2.3 Reverse transcriptase PCR (RT-PCR)

7.2.3.1 Cytochrome c

First-strand cDNA was synthesized using the Superscript kit (Amersham, UK). Approximately 5µg of CMC 40 leaf and NGA 1 storage root total RNA were used as templates, with either 1µg random hexamers or 1µg oligo $d(T)_{25}$ as primer and made up to 12µl with SMQ. After incubation at 70°C for 2min, the following components were added on ice: 1µl RNA guard, 4µl buffer, 1µl 0.1M DTT, 1µl 10mM dNTPs and 1µl Superscript enzyme (all from the kit; Amersham, UK). The reaction was incubated at 42°C for 1.5h, then 5µl used as a template for PCR of cytochrome c using the primers cyt1 and cyt2 listed above (Fig. 7.1).

PCR was carried out in a total reaction volume of 25µl containing the cDNA template, 1× NH_4 reaction buffer (16mM $(NH_4)_2SO_4$, 67mM Tris-HCl pH8.8, 0.01% (v/v) Tween-20; Bioline, UK), 3mM $MgCl_2$, 200µM dNTPs, 25pmol each forward and reverse primers and 1 unit *Taq* DNA polymerase (Bioline, UK), overlaid with mineral oil and amplified using a PTC-100 Programmable Thermal Controller (MJ Research Inc, USA). PCR conditions were 90s denaturation at 94°C, followed by 30 cycles of 94°C for 30s, 50°C for 30s, 72°C for 2min and a final extension at 72°C for 5min.

7.2.3.2 A41

RT-PCR was carried out for A41 with primers A and D using the same method as for cytochrome c.

To confirm the direction of A41, two different methods were used. The experiment involved using either primer A or D as the reverse-transcribing primer, with $dT_{17}N$ used as a control. Both Superscript (Amersham, UK) and *Tth Plus* DNA polymerase (Bioline, UK) enzymes were used. The protocol for Superscript was as above, while the latter employs a combined reverse transcription/polymerization assay.

For *Tth Plus*, approximately 100ng RNA template was used with 1× NH_4 reaction buffer, 1mM $MnCl_2$, 1μl RNasin, 250μM dNTPs, 20pmol reverse-transcribing primer (A, D or $dT_{17}N$), in a total volume of 19μl. These components were mixed, centrifuged briefly, then incubated at 65°C and allowed to cool to 20°C to anneal the primer. 1μl (4 units) of *Tth Plus* polymerase was added, mixed and the temperature raised slowly to 70°C using a PTC-100 programmable thermal controller (MJ research Inc., USA). The reaction was incubated at 70°C for 15min, placed on ice, then stopped by adding EGTA to a final concentration of 1.25mM.

PCR was carried out in the same tube with the following components added beneath the oil layer in a total reaction volume of 100μl. A final concentration of 1× chelating buffer (67mM Tris-HCl pH 8.8, 16.6mM $(NH_4)_2SO_4$, 0.75mM EGTA, 5% (v/v) glycerol, 0.01% (v/v) Tween-20) and 1.5mM $MgCl_2$ were used, together with 20pmol primer 2 (A or D). The PCR cycle was as follows: 94°C for 2min; 35 cycles of 94°C for 30s, 50°C for 30s, 72°C for 2min; 72°C for 5min and incubation at 4°C.

7.2.4 3' Rapid amplification of cDNA ends (3'-RACE)

First strand cDNA synthesized with Superscript enzyme was used for 3'-RACE (Frohman *et al.*, 1988). This uses PCR to amplify copies of the cDNA sequence between a selected position within the transcript and either the 5' or 3' end. A 25μl reaction was set up containing the following: 1× NH_4 buffer (as above), 1–3mM $MgCl_2$, 200μM dNTPs, 8pmol $dT_{17}G$, 5pmol primer A and 1 unit *Taq* DNA polymerase (Bioline, UK). The PCR cycle used was denaturation at 94°C for 2min; 40 cycles of 94°C for 30s, 55°C for 30s, 72°C for 2min; and a final extension at 72°C for 5min.

A second round amplification was done using primers A and $dT_{17}G$ on 1μl of first round template. Buffer components were as before but primer concentrations were doubled. Third round amplification was also set up using 1μl of second round

template. PCR products were electrophoresed on a 2% (w/v) NuSieve GTG agarose gel (FMC Bioproducts, USA) buffered in 1× TAE.

7.3 RESULTS

7.3.1 Sub-cloning of A41

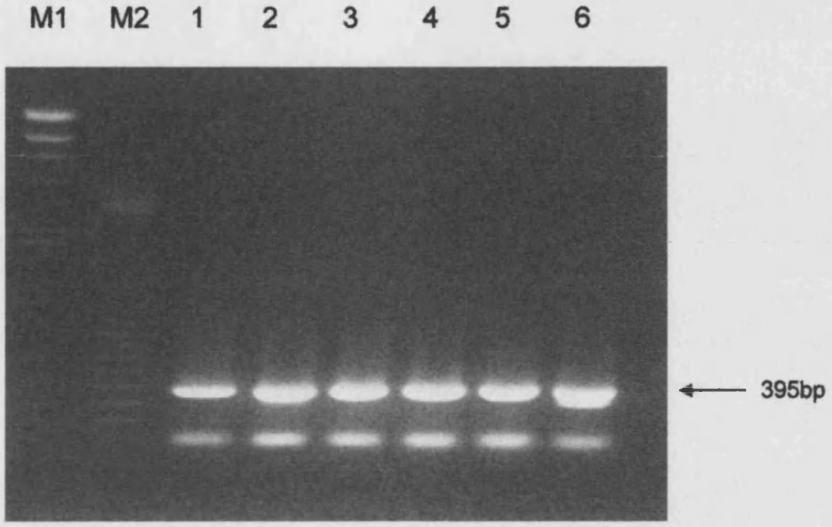
As with clones A24, A30 and A35, A41 was sub-cloned in to the pTAg vector (R&D Systems, UK). The PCR product was reamplified using the λgt11 forward and reverse primers on phage DNA obtained from a plate lysate, and electrophoresed. A clean product of the expected size was obtained from all reactions (Fig. 7.2). As usual, the DNA was purified away from the agarose, cleaned using a Sephadex column and approximately 50ng used for sub-cloning.

Blue-white selection enabled positive colonies to be identified and these were given a second check by single colony gel analysis, comparing them with the controls. The results from this can be seen in Figure 7.3.

7.3.2 Sequence analysis

Colonies 1, 2, 3, 5, 6, 15, 16, 17 and 19 appeared to be recombinants. DNA minipreps were done of these positives and one of each control, a small sample electrophoresed as usual and the same DNA migration pattern obtained (not shown). As before, a number of clones (here 3, 6 and 15) were sequenced manually in both directions for reliability.

Figure 7.2 PCR products pooled from five individual reactions (lanes 1–6) using 2μl phage DNA (plate lysate method) as template. Electrophoresis was at $5Vcm^{-1}$ in a 1% (w/v) agarose gel buffered with 1× TAE. Markers: M1: λ *Hin*dIII digest; M2: 100bp ladder.



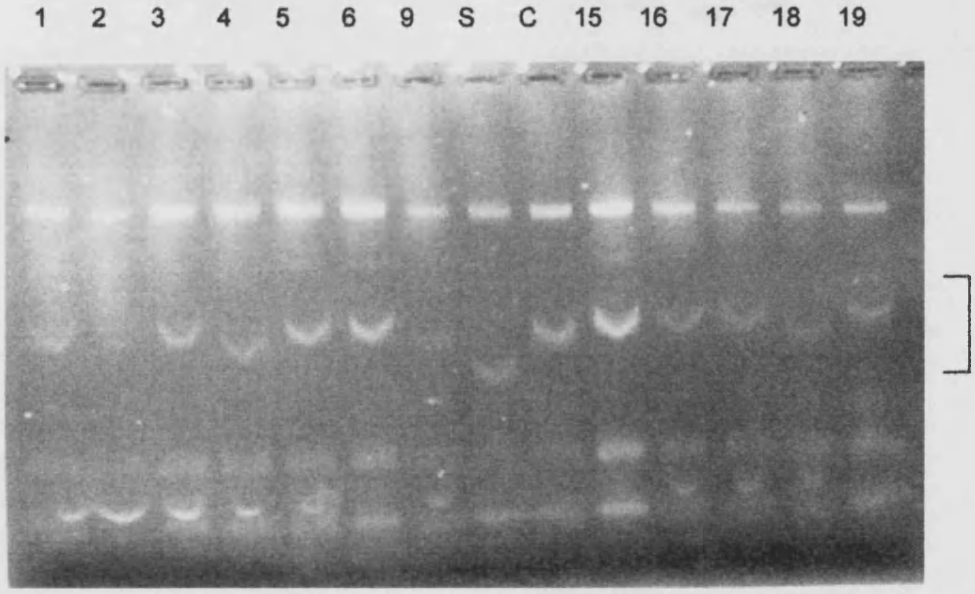


Figure 7.3 Single colony gel analysis of 12 white colonies of pTAg-A41, compared with the blue self-ligated control (S) and white control with 50bp insert (C). The bracket indicates the area to observe changes in DNA band size, compared with the standards.

These data were supplemented with information from automated sequencing reactions, also in both directions, to give the whole of the cDNA insert. Once the two cloning adaptors and flanking vector sequences had been removed, the size of A41 was 270bp. On translating the sequence in all six reading frames, a complete open reading frame of 88 amino acids was found to run throughout the cDNA in one frame only. This is shown below the nucleotide sequence in Figure 7.4.

Figure 7.5 shows the first five alignments from a BLAST (Altschul *et al.*, 1990) search to identify A41. The best comparisons were with human and yeast sequences but there was only limited homology over short stretches of A41. Thus a reliable indication as to its identity was not obtained with this fragment. However, the fact that there is an open reading frame over the entire sequence gave confidence that this perhaps represented part of a novel gene.

```
ATGTTCCATC CTGCAGTTGC CCACAAAGTA CGTGGTCCTG GCTTAAATTT
M  F  H  P   A  V  A   H  K  V   R  G  P  G   L  N  L

GCATAGAAGT TACATCCAAG AGAATTTTGA TGGAGAAGAG GTCAATGATA
 H  R  S   Y  I  Q  E   N  F  D   G  E  E   V  N  D  I

TTGAAAGAGA AGTATCTTCT CTAAAAGAGG ATTCAGATGA TATTGATGCA
  E  R  E   V  S  S   L  K  E  D   S  D  D   I  D  A

TTGTTGAGCT TGGAAGAGGA TGAACAGGAT GAATATGATG AGGAAGAAGT
L  L  S  L   E  E  D   E  Q  D   E  Y  D  E   E  E  V

TAGCACTGCA CGAACTTATG GAGATTATGG AAGCAATTCC CCTGATTCTT
 S  T  A   R  T  Y  G   D  Y  G   S  N  S   P  D  S  Y

ATTCGACTTG TGGTTCTAAA
  S  T  C   G  S  K
```

Figure 7.4 Nucleotide sequence of clone A41, showing the possible translation below. Data was obtained from sequencing reactions along both strands of the cDNA, using either the purified PCR product or DNA from sub-cloned A41 as template.

Figure 7.5 BLAST results for the 270bp cassava cDNA within clone A41, showing the first five nucleotide sequence comparisons. Seq 1: *H. sapiens* partial cDNA sequence 1; Seq 2: *H. sapiens* partial cDNA sequence 2; Seq 3: *S. cerevisiae* MFT1 genomic DNA sequence; Seq 4: *S. cerevisiae* PLC2 gene involved in mitochondrion protein import; Seq 5: *X. laevis* cDNA encoding hepatocyte nuclear factor 4.

```
A41:      22 CTGGCTTAAATTTGCATAGAAGTTACATCCAAGAGAATTTTGATGGAGAAGAGGTCAATGATATT
             || |||||||||||| | | | |||    | | | | ||| | |||||||  || |    || ||
Seq 1:   230 CTTGCTTAAATTTGCTTCGGAATTAGGCACCATAAACTTTAGTTGGAGAATGGGACTTAAATTTT

A41:         GAAAGAGAAGTAT   99
             | | | | ||||
Seq 1:       CACATATATGTAT  307



A41:      22 CTGGCTTAAATTTGCATAGAAGTTACATCCAAGAGAATTTTGATGGAGAAGAGGTCAATGATATT
             || |||||||||||| | | | |||    | | | | ||| | |||||||  || |    || ||
Seq 2:   230 CTTGCTTAAATTTGCTTCGGAATTAGGCACCATAAACTTTAGTTGGAGAATGGGACTTAAATTTT

A41:         GAAAGAGAAGTAT  99
             | | | | ||||
Seq 2:       CACATATATGTAT 307



A41:      36 CATAGAAGTTACATCCAAGAGAATTTTGATGGAGAAGAGGTCAATGATATTGAAAGAGAAGTATC
             |||  |  | ||   || ||| ||   ||    |||||||   ||||||| |||  ||| ||
Seq 3:   962 CATGAATATGACGATCAGGAGGATGAAGAAAATGAAGAGGAAGATGATATGGAAGTAGACGTTGA

A41:         TTCTCTAAAAGAGGATTCAGATGATATTG  129
                  ||||||||||||   || | |  ||
Seq 3:       GGATATAAAAGAGGATAATGAAGTTGATG 1055



A41:     129 CAATATCATCTGAATCCTCTTTTAGAGAAGATACTTCTCTTTCAATATCATTGACCTCTTCTCC
             ||  | | ||   ||||||||||| |      || |||  ||| |||||||   |||||||
Seq 4: 2069 CATCAACTTCATTATCCTCTTTTATATCCTCAACGTCTACTTCCATATCATCTTCCTCTTCATT

A41:         ATCAAAATTCTCTTGGATGTAACTTCTATG    36
             ||   || ||| ||   || |  |  |||
Seq 4:       TTCTTCATCCTCCTGATCGTCATATTCATG  2162



A41:      70 AAGAGGTCAATGATATTGAAAGAGAAGTATCTTCTCTAAAA 110
             |||||||||||||| || ||||| |  |  ||||||  |||
Seq 5:   765 AAGAGGTCAATGATGTTTAAAGACATATTGCTTCTCGGAAA 805
```

7.3.3 Southern analysis

A41 was used for Southern analysis, to verify that it was from cassava. Figure 7.6 shows the *Eco*RI and *Hind*III digestion patterns on the two leaf gDNA samples, from cultivars CMC 40 and MCol 1684. Hybridisation of ^{32}P-labelled A41 to these digests is shown alongside. Neither of the two enzymes used had a restriction site within A41, to make analysis simpler. The autoradiographs show a single band in each lane for the cassava digests only, indicating that A41 is a cassava sequence. Since the original level of hybridization was very intense (B), particularly for the CMC 40 digests, a second autoradiograph with a shorter exposure time was done (C). As only one band was obtained from the restriction digests, the gene corresponding to A41 may be present as a single copy within the genome. Neither site was present within the coding region sequenced so far from the cDNA and sites within the non-coding region are less likely to be conserved between genes.

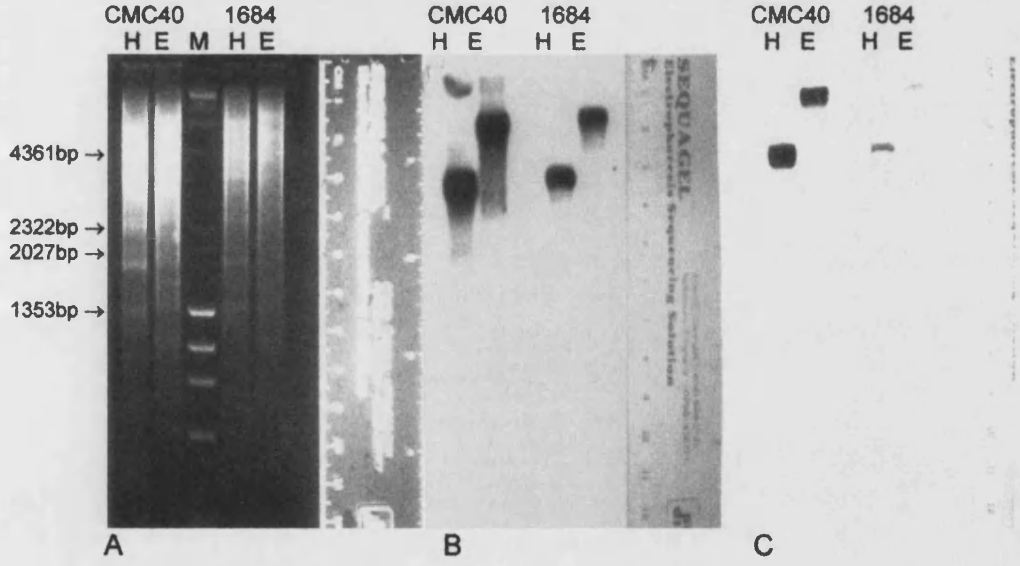
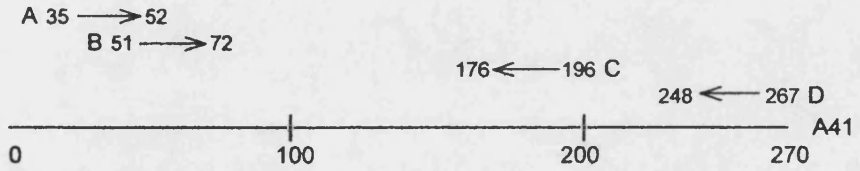


Figure 7.6 A: Single digests of 10μg CMC 40 and 5μg MCol 1684 gDNA with *Eco*RI (E) and *Hind*III (H), with λ DNA *Hind*III/ϕX-174-*Hae*III markers (M). Electrophoresis was in a 1% (w/v) agarose gel, at 1.5Vcm^{-1} overnight. The autoradiographs of the Southern blot probed with A41 are shown to the right: B: developed at –70°C for 7d, C: developed at –70°C overnight.

7.3.4 Extension of A41

The Southern results were very encouraging. As A41 appears to be a single copy gene, it is not so surprising that the small fragment identified by PCR showed little reliable homology with any of the database accessions. It was decided to try and extend the cDNA sequence information to identify A41, before using it to screen a sample of a genomic library and the technique of 3'-RACE was chosen (as described in **7.2.4**). As the aim was merely to extend A41 at this stage, it was decided to amplify the 3' end of the fragment, using an oligomer of dT and a primer internal to A41, since this is simpler than getting the 5' end. Primers internal to A41 were therefore needed, and designed using the PRIME (GCG) programme.

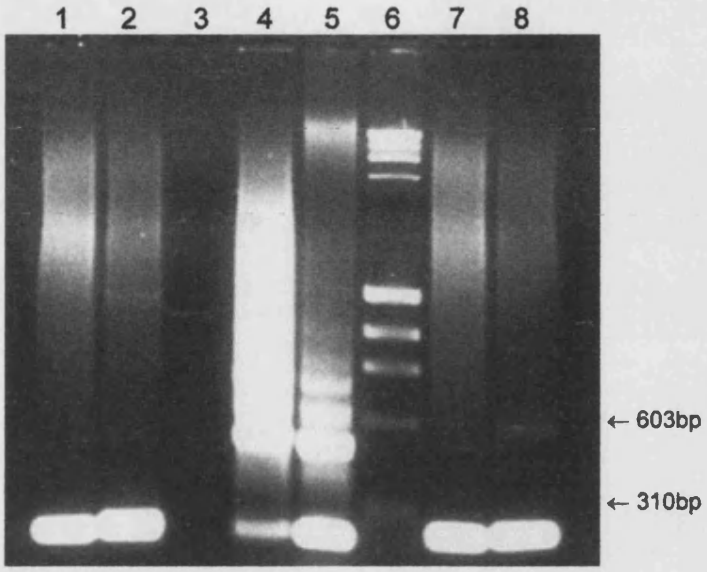
Of the various primers suggested by the computer programme, two forward and two reverse primers, with positions and orientation as shown in Figure 7.7, were selected. These were tested to ensure the correct size product was amplified (Figure 7.8).



A–D: 232bp
B–D: 216bp

Figure 7.7 Diagram of the orientations and positions within the A41 sequence of primers A, B, C and D.

Figure 7.8 Amplifications in 1% (w/v) agarose at $5Vcm^{-1}$ using the A41 internal primers on the following template DNA. Lane 1: B–D on CMC 40 gDNA; 2: A–D on MCol 1684 gDNA; 3: negative control; 4: B–D on A41 PCR product; 5: B–D on A41 phage DNA; 6: λ *Hin*dIII//ϕX-174-*Hae*III molecular weight markers; 7: A–D on CMC 40 gDNA; 8: B–D on MCol 1684 gDNA.



7.3.4.1 RT-PCR

To determine the orientation of A41, and hence decide which primer internal to A41 was required for amplification during 3'-RACE, RT-PCR was done. Since an open reading frame of 90 amino acids ran through the sequence, it was likely that this was also the 5'–3' direction, but this needed to be confirmed. A positive control for this was to use the amplification of cytochrome c. Degenerate primers designed by Dr A Bailey to amplify the sequence from the fungus *Metarrhizium anisopliae* used a number of sequences for cytochrome c from a variety of organisms, one of which was *A. thaliana*. Samples of these primers were kindly donated for the cassava experiment with the hope that the cassava sequence would be sufficiently similar to the *Arabidopsis* sequence to allow amplification.

Figure 7.9A shows the PCR product for cytochrome c from both of the genomic DNA preparations in lanes 1 and 2, with a major band adjacent to the marker fragment of 1078bp. *S. cerevisiae*, used as a control, does not have an intron within the genomic sequence, and a major product of the expected size of 219bp can be seen in lane 4. Since *A. thaliana* has an intron present of around 1kb, the cassava product was expected to be over 1kb in size.

The corresponding RT-PCR product for the cassava cytochrome c is shown in Figure 7.9B. This was amplified from leaf total RNA, using random primers for the first strand cDNA synthesis, and was around the expected cDNA size of 219bp.

The A–D A41 RT-PCR product of 232bp is shown in Figure 7.9C. This was again amplified from the leaf RNA sample using hexamer primers. No amplification was obtained using the storage root RNA preparation for either of these products (data not shown). Since A41 was isolated from mRNA and cytochrome c is expected to be ubiquitously expressed, this was assumed to be due to degradation of the RNA sample.

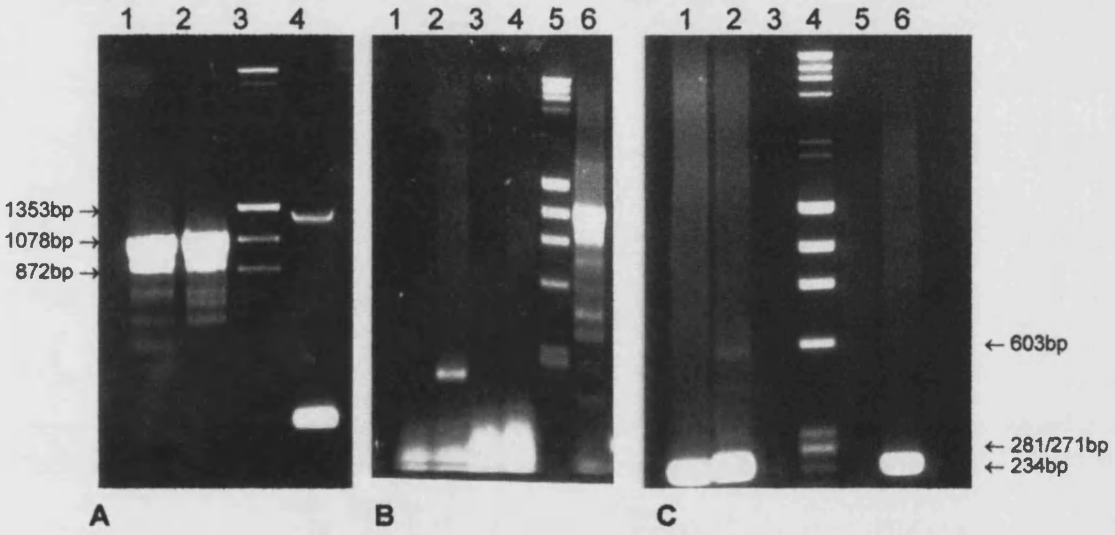
Finally, Figure 7.10 shows the A–D RT-PCR product amplified from leaf RNA samples using primers a and d separately as reverse-transcribing primers.

7.3.4.2 3'-RACE

RT-PCR had been carried out successfully for both cytochrome c and A41. Subsequently, amplifications were done using anchored dT oligomers. Again, cytochrome c was used as the positive control, with the oligomer $dT_{17}G$. Successful A41 amplification using A–D on 1st strand cDNA reverse-transcribed using $dT_{17}G$, was also achieved (Fig. 7.11), therefore $dT_{17}G$ was used for the 3'-RACE.

Figure 7.9 **A**: PCR of cytochrome c; **B**: RT-PCR amplification of cytochrome c from cassava; **C**: A41 RT-PCR.
A: Lane 1: CMC 40 gDNA template; 2: MCol 1684 gDNA template; 3: λ *Hin*dIII/ϕX-174-*Hae*III digest (Pharmacia, USA); **4**: *S. cerevisiae* gDNA.
B: Lane 1: CMC 40 leaf RNA using $d(T)_{25}$ as R-T primer; 2: CMC 40 leaf RNA using random hexamers as R-T primers; 3: NGA1 storage root RNA using $d(T)_{25}$ as R-T primer; 4: NGA1 storage root RNA using random hexamers as R-T primers; 5: λ *Hin*dIII/ϕX-174-*Hae*III digest markers (Pharmacia, USA); 6: control amplification of cytochrome C from CMC 40 leaf gDNA.
C: Lane 1: A–D control PCR on CMC 40 leaf gDNA template; 2: A–D control PCR on MCol 1684 leaf gDNA template; 3: negative control with no DNA; **4**: λ *Hin*dIII/ϕX-174-*Hae*III digest markers (Pharmacia, USA); 5: A–D RT-PCR on CMC40 storage root RNA; 6: A–D RT-PCR on NGA1 leaf RNA. Both RT-PCR reactions used random hexamers to generate the first-strand cDNA.

All PCR products were electrophoresed in 1% (w/v) agarose and run at $5Vcm^{-1}$. Appropriate bands are indicated in bp.



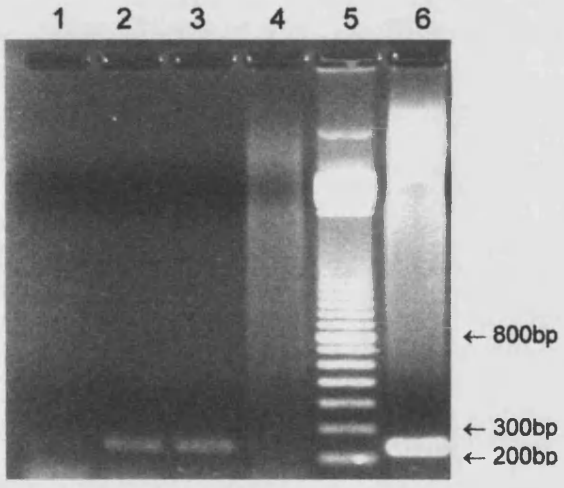
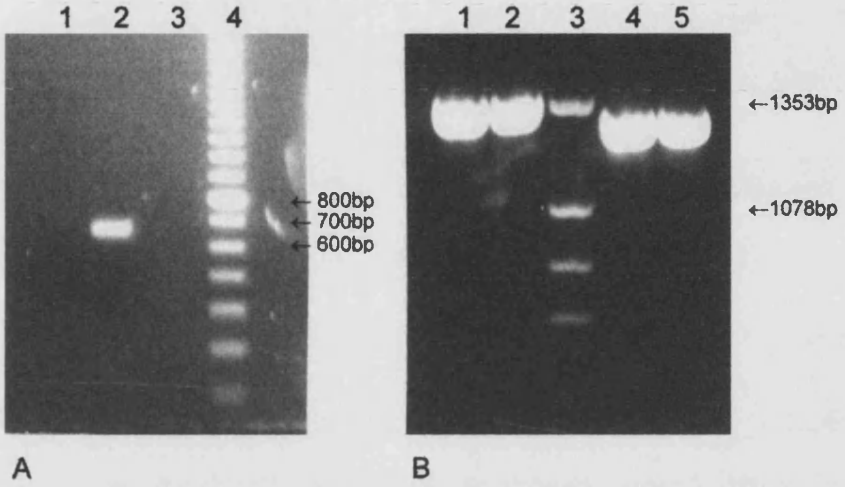


Figure 7.10 Electrophoresis at $5Vcm^{-1}$ in 1% (w/v) agarose of RT-PCR products on two leaf RNA samples A and B. Lanes 1–3: amplification of A–D product using (1) primer a; (2) primer d; and (3) dTN as the reverse-transcribing primer on RNA from leaf sample A respectively. Lane 4: A–D product from dTN RT primer on leaf sample B RNA. (Lanes 1–4 using *Tth* Pol enzyme). Lane 5: 100bp ladder (Pharmacia, USA). Lane 6: A–D product from dTN RT primer on leaf sample A RNA.

Figure 7.11 **A**: Amplifications using $dT_{17}G$ and primer A on first round templates buffered with (lane 1) 1mM $MgCl_2$, (lane 2) 2mM $MgCl_2$. Lane 3 has the negative control and lane 4 shows 100bp ladder DNA markers (Pharmacia, USA). Electrophoresis in 1% (w/v) agarose, 1× TBE. **B**: Third round amplifications. Lanes 1 and 2: 3'RACE using 1mM $MgCl_2$ second round template; Lane 3: λ *Hin*dIII/ϕX-174-*Hae*III molecular weight markers (Pharmacia, USA); Lanes 4 and 5: 3'-RACE using 2mM $MgCl_2$ second round template. Electrophoresis in 2% (w/v) NuSieve agarose (FMC BioProducts, USA) buffered in 1× TAE.



7.3.4.3 Sub-cloning of 3' RACE product

A single band of approximately 650bp was obtained for the second round amplification from the 2mM $MgCl_2$ sample (Figure 7.11A). Third round amplifications used 1μl of each of the 1mM and 2mM $MgCl_2$ second round reactions and the results can be seen in Figure 7.11B. It was assumed that the large quantities of DNA loaded, combined with the high percentage gel, gave anomalously high apparent molecular weights. Since a minor product of higher molecular weight was obtained in the amplifications using the 2mM $MgCl_2$ sample (results not shown), those from the 1mM $MgCl_2$ second round were used for sub-cloning.

These bands were excised from the gel, purified by electroelution and a small sample run on a 1% (w/v) agarose gel to check the quality. The approximate molecular weight this time was 500bp, which corresponds better with the original second round amplification estimate. Sub-cloning was carried out using 10–20ng cDNA and the

pUAg vector (R&D Systems, UK). The transformation results were poor, but some positives were obtained. DNA was prepared by the miniprep method, using a blue colony from the negative control reaction (no DNA) for comparison.

7.3.4.4 DNA Sequencing

Multiple clones of pUAg-A41 were sequenced. Approximately 670bp were obtained using primer B, which included most of the original fragment. Primers located within the pUAg vector (R&D Systems, UK) were used to obtain sequence at the 5' end of the A41 3'-RACE product and also sequence in the opposite direction. Figure 7.12 shows this extended cDNA sequence together with the continuation of the open reading frame of 166 amino acids. BLAST searches for the whole of A41 showed limited homology. However, by using the SMART programme (Schultz *et al.*, 1998) to identify possible domains within the protein, a coil was found at residues 29–62 and a helix-loop-helix region at residues 109–157. A search was conducted to find proteins with similar domains (although additional domains may be present in these proteins) and numerous transcription factors were listed. When the BLAST and SMART results were compared with each other, the similarities between proteins were found generally within the residues comprising the helix-loop-helix, the region that binds to DNA and is characteristic of this class of transcription factors. Some of these results can be seen in Figure 7.13.

Figure 7.12 cDNA sequence obtained from multiple clones of pUAg-A41 containing the 3'-RACE product, with translated open reading frame below. TAA: stop codon; underlined: putative coil region from residues 29–62; bold: helix-loop-helix region.

```
ATGTTCCATC CTGCAGTTGC CCACAAAGTA CGTGGTCCTG GCTTAAATTT
M  F  H  P   A  V  A   H  K  V   R  G  P  G   L  N  L

GCATAGAAGT TACATCCAAG AGAATTTTGA TGGAGAAGAG GTCAATGATA
 H  R  S   Y  I  Q  E   N  F  D   G  E  E   V  N  D  I

TTGAAAGAGA AGTATCTTCT CTAAAAGAGG ATTCAGATGA TATTGATGCA
  E  R  E   V  S  S   L  K  E  D   S  D  D   I  D  A

TTGTTGAGCT TGGAAGAGGA TGAACAGGAT GAATATGATG AGGAAGAAGT
L  L  S  L   E  E  D   E  Q  D   E  Y  D  E   E  E  V

TAGCACTGCA CGAACTTATG GAGATTATGG AAGCAATTCC CCTGATTCTT
 S  T  A   R  T  Y  G   D  Y  G   S  N  S   P  D  S  Y

ATTCGACTTG TGGTTCTAAA CCCAGGAAAA ATGGATCTTC CTCTATTCAG
  S  T  C   G  S  K   P  R  K  N   G  S  S   S  I  Q

AGGTTCTCTG GAAATGATAG CAGCTGTAAC AGTGAAAGGA AACGGCAGAA
R  F  S  G   N  D  S   S  C  N   S  E  R  K   R  Q  K

AATGAAGAAG ATGGTGAAGG CACTGAGAGG AATTGTACCT GGTGGTGACC
 M  K  K   M  V  K  A   L  R  G   I  V  P   G  G  D  Q

AAATGAATAC GGTAACAGTT CTTGATGAAG CCGTCAGATA TCTGAAGTCT
  M  N  T   V  T  V   L  D  E  A   V  R  Y   L  K  S

CTCAAAGTTG AAGTGCAGAA GATTGGAGTT GGGAATTTTA AGAATTAAAT
L  K  V  E   V  Q  K   I  G  V   G  N  F  K   N  *

TTTAATTTCA TCTGATATTT AACTCCATTC TCTAAGATAC AGGCAGTTCT

GCTGTCTGCA TCAATTAGTT TCCCTCTCTC TTGGCTGGTT GGCATTAGGT

GAATAATTCT GACCGGTGTT GGGCACCCAT GTTCATGATC CAACGATCCA

ATTATAGATT GGTTGCCACA AAATGAAGAC CCTAGTGGTG GTCTCCTCTT

AAAATTTTCT GTAATTTTAG TTTCTACGCA AATTTAAGCC AGGACA
```

Figure 7.13 BLAST results comparing the translated A41 sequence with other amino acid sequences for various DNA-binding proteins. A: *Arabidopsis thaliana* sequence with similarity to transcriptional activator Ra; B: Maize SN mRNA, involved in tissue-specific anthocyanin expression; C: *Arabidopsis thaliana* phytochrome-associated protein 3; D: *Pennisetum glaucum* myc-like regulatory R gene product; E: *Arabidopsis thaliana* sequence similar to symbiotic ammonium transporter, SAT1; F: *Arabidopsis thaliana* sequence for putative phaseolin G-box binding protein.

```
A41:  101 RFSGNDSSCNSERKRQKMKKMVKALRGIVPGGDQMNTVTVLDEAVRYLKSLKVEVQ 156
          R S +   +   + R+R+++ + ++ L+ IVPGG +M+T ++LDEA+RY K LK +V+
A:    117 RISDDPQTVVARRRRERISEKIRILKRIVPGGAKMDTASMLDEAIRYTKFLKRQVR 172


A41:   83 SYSTCGSKPRKNGSSSIQRFSGNDSSCNSERKR-QKMKKMVKALRGIVPGGDQMNT 141
          ++ +CG      G++      +G       SERKR +K+ +M    L+ ++P    ++N
B:    401 AWESCGGA---TGAAQEMSGTGTKKHVMSERKRREKLNEMFLVLKSLLPSIHRVNK 457

A41:  142 VTVLDEAVRYLKSLKVEVQKI 158
          ++L E + YLK L+  VQ++
B:    458 ASILAETIAYLKELQRRVQEL 474


A41: 75 DYGSNSPDSYSTCGSKPRKNGSSSIQRFSGNDSSCNSERKRQKMKKMVKALRGIVP 130
         D    S D     G  P + G  S +  S    + +  R+R ++ + ++AL+ ++P
C:  318 DVEEESGDGRKEAG--PSRTGLGSKRSRSAEVHNLSERRRRDRINEKMRALQELIP 371


A41:131 GGDQMNTVTVLDEAVRYLKSLKVEVQ 156
         ++++   ++LDEA+ YLKSL+++VQ
C:  372 NCNKVDKASMLDEAIEYLKSLQLQVQ 397


A41: 111 SERKR-QKMKKMVKALRGIVPGGDQMNTVTVLDEAVRYLKSLKVEVQKI 158
         SERKR +K+ +M  AL+ +VP   +++  ++L E + YLK L+  VQ++
D:    6 SERKRREKLNEMFLALKSLVPSIHKVDKASILAETIAYLKELQRRVQEL 54


A41:   111 SERKR-QKMKKMVKALRGIVPGGDQMNTVTVLDEAVRYLKSLKVEVQKI 158
           +ERKR QK+ + + AL  ++PG  + +  TVL++A+++LK L+  V+K+
E:     136 AERKRRQKLNERLIALSALLPGLKKTDKATVLEDAIKHLKQLQERVKKL 184


A41:    78 SNSPDSYSTCGSKPRKNGSSSIQRFSGNDSSCN 110
           S+SP S S+ GS+     +S +  F   D+  N
E:      66 SSSPSSSSSSGSR-----TSQVISFGSPDTKTN 93


A41:   111 SERKRQ-KMKKMVKALRGIVPGGDQMNTVTVLDEAVRYLKSLKVEVQKI 158
           SER+R+ +MK  + ALR +VP   +M+  +++ +AV Y++ L+ + +K+
F:     135 SERRRRGRMKDKLYALRSLVPNITKMDKASIVGDAVLYVQELQSQAKKL 183


A41:    75 DYGSNSPDSY-STCGSKPRKNGSSSIQRFS-GNDSSCNS 111
           D  SN  +S+ S  G   R  G    + ++ G+DSS  +
F:      78 DLDSNVAESFRSFDGDSVRAGGEEDEEDYNDGDDSSATT 116


A41:    16 NLHRSYIQENFDGEEVNDIEREVSSLKEDSDDIDA 50
           N+  S+   +FDG+ V     E      D DD  A
F:      82 NVAESF--RSFDGDSVRAGGEEDEEDYNDGDDSSA 114
```

7.3.5 Analysis of the genomic clone corresponding to the A41 cDNA

7.3.5.1 Screening of the cassava genomic library

The first round library screen using 50ng A41 labelled with around 25μCi α-^{32}P-dCTP gave good results, with four positives identified as expected from the 100000 plaques screened using duplicate membranes. After purification by another two rounds of screening, several clones for each original positive were obtained. DNA from a number of these was then used for sequencing.

7.3.5.2 Sequencing of individual gA41 clones

Primers b and d from the sequencing of the A41 cDNA product (see Table 7.1) were chosen for the first two sequencing reactions of the genomic clone. Positive results of approximately 600bp for each confirmed that the clone was indeed that corresponding to A41, and positioned the cDNA relative to the genomic sequence. This new sequence was then used to design primers for the next set of reactions, using the PRIME (GCG) programme. Table 7.2 lists the primers used while Figure 7.14 shows the relative extent, position and direction of each reaction. Almost 3.3kb were eventually sequenced, most of this in both directions.

Table 7.2 Primers used in the sequencing of the A41 genomic clone

Primer name	Sequence 5'–3'
Forward primers:	
gf1	GCA TAG AAG TTA CAT CCA AGA G
gf2	GCA GAA GAT TGG AGT TGG
gf3	GGC CTG CTA CTA TTT AG
gf4	AGT CTC CAT CAC TTG CCA CA
gf5	GTG TGG AGG TCC TCA AGG AT
gf6	GCT GAA CTT AGA CCG TTA CC
Reverse primers:	
gr1	AGA ACC ACA AGT CGA ATA AG
gr2	GGT GCA ACT GGA ATA TG
gr3	TCC GTC CCA AAA CCC AAT C
gr4	CAG CAC CAA GCC AGT ACC AAA G
gr5	GCC TTC AGG TAA CGG TCT AAG TTC
gr6	GGC AGT AAC AGC ATT AGA TCG TCT C

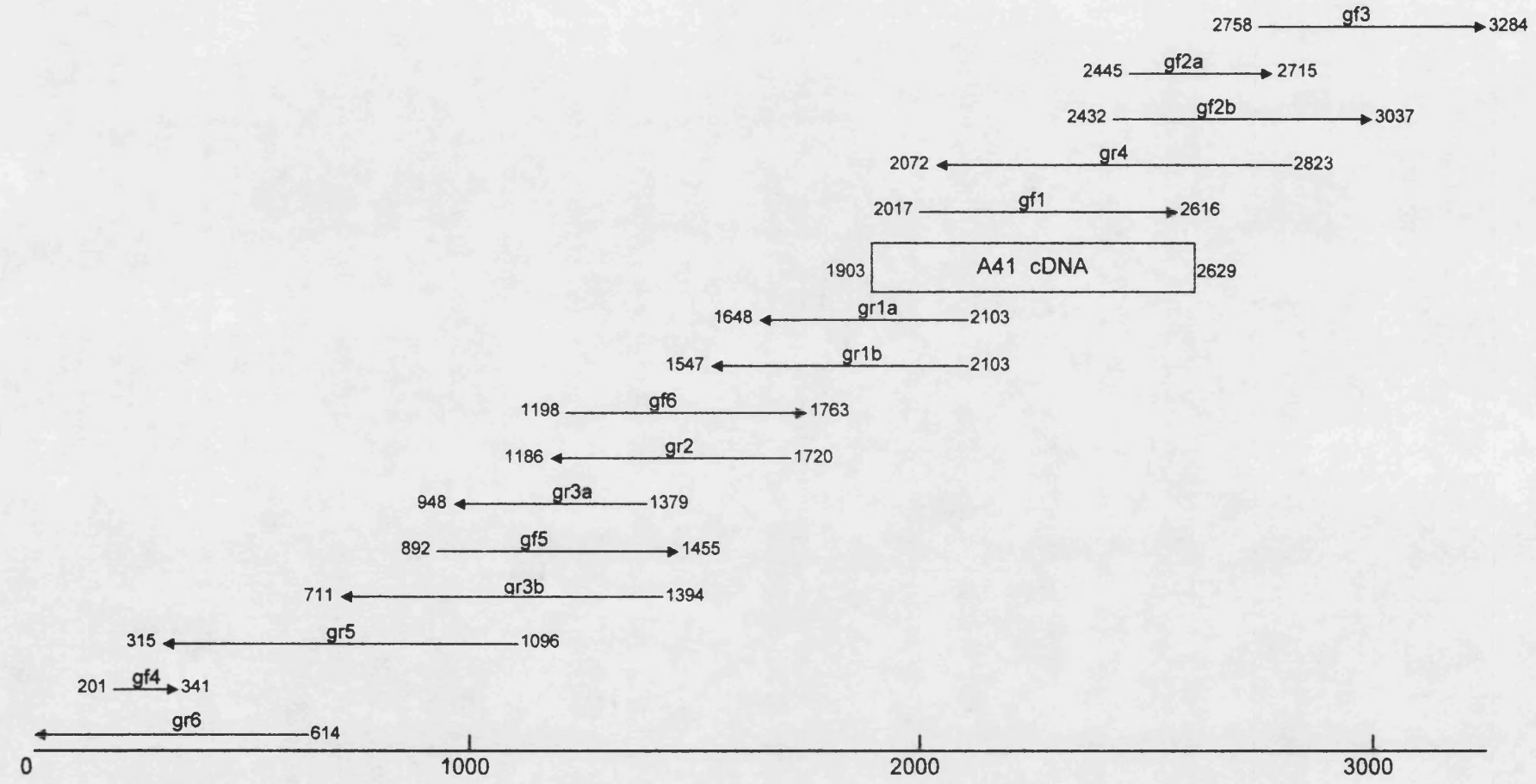


Figure 7.14 Diagram showing the various sequencing runs covering the genomic sequence of A41. The name of each reaction corresponds to the primer used; some primers were used more than once, indicated a and b. Direction is indicated by the arrowheads and the position of the cDNA sequence is also shown.

The open reading frame obtained with the cDNA sequence was extended, by 76 residues in the 5' direction. In an attempt to identify A41 conclusively, further BLASTN and BLASTX searches were done using both this translated sequence and the full genomic sequence (Altschul *et al.*, 1990, 1997). Again, the same sections within the putative helix-loop-helix region indicated in Figure 7.11 were recognized, being compared with the same proteins as before (Figure 7.12).

When the full amino acid sequence for the *Arabidopsis* putative transcription factor *Ra* was aligned with the continuous open reading frame from cassava, only the areas indicated by the computer showed any degree of similarity (Figure 7.15).

The full genomic sequence obtained for A41 is given in Figure 7.16. The GENSCAN programme (version 1.0; Burge and Karlin, 1997), which identifies gene structures, predicted that the open reading frame obtained corresponds to a terminal exon, rather than representing a single exon gene, with the probable 3' splice site as shown and a polyadenylation signal at position 2731–2736. However, the programme failed to identify the corresponding 5' splice site, an initial exon or a translation start site. Nucleotides 89–96 were proposed as the transcription start site (TCS).

```
Me  GLLMHSDQQFHPKEVPHFANQRGNSYMHIPVAPSFAAALPPTGKHLMPVHGIEFQPSEIC

At  MENGMYKKKGVCDSCVSSKSRSNHSPKRSMMEPQPHHLLMDWNKANDLLTQEHAAFLNDP
Me  PKNFIIFDQNDHRSQVMFHPAVAHKVRGPGLNLHRSYIQENFDGEEVNDIEREVSSLKED

At  HHLMLDPPPETLIHLDEDEEYDEDMDAMKEMQYMIAVMQPVDIDPATVPKPNRRNVRISD
Me  SDDIDALLSLEEDEQDEYDEEEVSTARTYGDYGSNSPDSYSTCGSKPRKNGSSSIQRFSG

At  DPQTVVARRRRERISEKIRILKRIVPGGAKMDTASMLDEAIRYTKFLKRQVRILQPHSQI
Me  NDSSCNSERKRQKMKKMVKALRGIVPGGDQMNTVTVLDEAVRYLKSLKVEVQKIGVGNFK

At  GAPMANPSYLCYYHNSQP
Me  N
```

Figure 7.15 Alignment between the putative transcription factor *Ra* from *Arabidopsis thaliana* (At) and the translated region of the genomic sequence of A41 (Me; 1668bp–2393bp). Identical residues are indicated by dashes.

Figure 7.16 Genomic sequence obtained for A41 (3670bp). Possible control regions and transcription starts are highlighted in blue, and translation starts are highlighted in yellow. The various potential 5' splice sites corresponding to an exon-intron boundary are shown in green, as is the predicted 3' splice junction. Vertical dotted lines correspond to the end of the reading frame as expected if these sites represent genuine boundaries, and vertical arrows represent the splice points themselves. Two possible polyadenylation signals are underlined and two almost identical inverted repeats are overlined. Where relevant, translations of the nucleotide sequence are underneath the DNA, the single-letter amino acid code placed below the first nucleotide of the codon.

Figure 7.16 Genomic sequence obtained for A41 (continued).

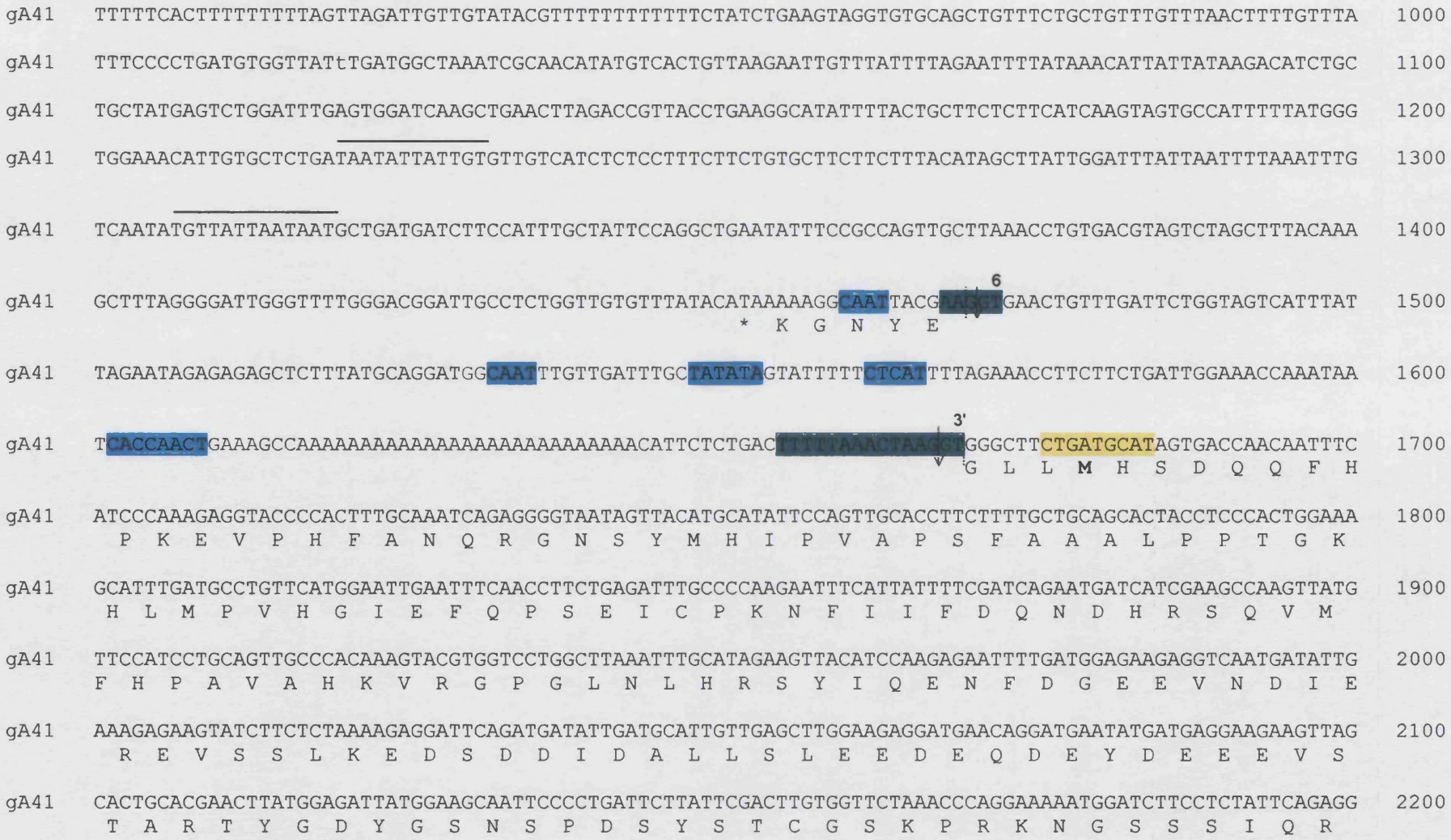


Figure 7.16 Genomic sequence obtained for A41 (continued).

```
gA41   TTCTCTGGAAATGATAGCAGCTGTAACAGTGAAAGGAAACGGCAGAAAATGAAGAAGATGGTGAAGGCACTGAGAGGAATTGTACCTGGTGGTGACCAAA   2300
       F  S  G  N  D  S  S  C  N  S  E  R  K  R  Q  K  M  K  K  M  V  K  A  L  R  G  I  V  P  G  G  D  Q  M

gA41   TGAATACGGTAACAGTTCTTGATGAAGCCGTCAGATATCTGAAGTCTCTCAAAGTTGAAGTGCAGAAGATTGGAGTTGGGAATTTTAAGAATTAAATTTT   2400
         N  T  V  T  V  L  D  E  A  V  R  Y  L  K  S  L  K  V  E  V  Q  K  I  G  V  G  N  F  K  N  *

gA41   AATTTCATCTGATATTTAACTCCATTCTCTAAGATACAGGCAGTTCTGCTGTCTGCATCAATTAGTTTCCCTCTCTCTTGGCTGGTTGGCATTAGGTGAA   2500

gA41   TAATTCTGACCGGTGTTGGGCACCCATGTTCATGATCCAACGATCCAATTATAGATTGGTTGCCACAAAATGAAGACCCTAGTGGTGGTCTCCTCTTAAA   2600

gA41   ATTTTCTGTAATTTTAGTTTCTACGTACTCTTTAGTGGAGTGAAGATTGCCTTGCCATTCAATATGTACTACCCTCTTCTACTAGCAGGTATTCTGTCTG   2700

gA41   GCCTGCTACTATTTAGTACTCCTTTTAAATAATGCTATGTTGTACTGTACACTTTTATGTCTATTTAACTGTTCCCATTTCCATTGCACTCAAGCAGTGG   2800

gA41   TTTATATGTTCTTGTCTGTCTTTGGCATGTGCTCATTATCAATGGTCCTTTGGTACTGGCTTGGTGCTGGATCCAGTTGTGCTTCGTTAAATGCTTAAAC   2900

gA41   TTAAATGGGCTTGCTTTGTGTCAGCATGTGTTACTGAAAAATCTGCTACTCTACTTTTTGTTTTTGAAAGATGCTTCAAACATACTTGTGGTTTCCCTTT   3000

gA41   ATCCTCTAATGCAATTGTAAATACTTATTGTGCTGAGAATTTAGAGACATGAAGATGTTGATGATGCACTTATTGTGGTTTGGCACCTTTGTTGCAGTCC   3100

gA41   CTGTATGTAGTGGTGTCAGTGGAGCACTATAGCAGTATGAGCATCTTTTTATGTAATGTTGGATCCAAGTTGTATTCTTTTTATGGGAACCCTAGTGGGT   3200

gA41   AGTCCATTACTTTATGTGAGCATGACTAGCTCAGATCTCTTTCAATGTCTTTATCCATCTTTTAAATATTTTTTTATATATGTGCCCTTTTATCCTCCTT   3300

gA41   TATGTGCTGGAATTCTAGGGACATGGAATCAATGTCATGCATGCATGCATGCCCAGTGCCTGCCTTTTTTTTTTATTATTTTCTCTATTTTATAATTTTG   3400

gA41   CATAATTCACATACTATTTGCAGGAGAAATTATATATAGATGAGTAATCTCATTATATTGTAGGATGTATATAATGCTTGATAGCATTTAAGGaCTGGGG   3500

gA41   GAGGGGAAGTATTGAATTTGAGACCCCAAGATCAAGCTTCATGTTCTTACCACCAAACTATGAGATATGTTAAATGTATAAAAAGTTATTTTGCTCTTTT   3600

gA41   AATTAAATTATAAATTTAATATTTATAATTTAAAAATAAAAGAGATAAATATGATAAAAGATAAAGTTGA                                   3670
```

7.4 DISCUSSION

7.4.1 Coding regions of gA41

The genomic clone corresponding to the A41 cDNA isolated from the original storage root library was difficult to characterize. Computer analysis indicated only positions for a terminal exon with 3' splice site but no corresponding 5' site. Evidence for the terminal exon is fairly good – there is a polyadenylation signal (underlined) 331bp after the stop codon of the ORF. The 3' splice site in dicotyledons is the sequence $T_{10}G(C/T)AG\downarrow GT$, with the AG preceding the splice site almost invariant in naturally occurring higher plant introns (Simpson and Filipowicz, 1996). In A41, this sequence (located at 1659–1673bp and indicated in Figure 7.16) reads $T_5AAACTAAG\downarrow GT$, which is a reasonable agreement. However, when the entire genomic sequence was examined by eye, it was difficult to confidently predict a 5' splice site. Intron size in plants varies, with both monocots and dicots capable of processing introns of several kb, though many introns are much shorter than this, being between 80 and 100bp in length (reviewed in Hughes, 1996). The consensus sequence in dicots is $AAG\downarrow GTAAGT$, with the first two nucleotides of the intron (GT) again almost invariant. Six possible 5' splice sites were found: $AAG\downarrow GT$, located at 196–200bp upstream of the actual splice point of the 3' site; $AG\downarrow GT$ at 721–724bp, $AG\downarrow GT$ at 725–728bp; $AG\downarrow GT$ at 837–840bp; $AAG\downarrow GT$ at 503–507bp; and $AG\downarrow GT$ at 417–420bp (all highlighted in Figure 7.16). These 5' sites could mark the end either of the initial exon of A41 or of a subsequent exon. For the first possibility, a translation start site is required upstream of the splice site, with continuous open reading frame to the starting methionine. For the second, a further 3' site would be needed, again in frame. A translation is shown below the A41 sequence for the six possible sites. The first has an open reading frame for some distance, but there is no

methionine residue within it, while the second and third both have a stop codon as the second upstream residue. The fourth and fifth sites both have a continuous reading frame which includes a methionine residue. However, the transcription start site (TCS), a CTCATCA consensus in plants, and the TATA and CAAT/AGGA boxes which would be expected upstream do not appear to be present, and no evidence for another 3' site was found. A final objection to either of these sites is the base composition of the proposed exon, in both cases very AT-rich which is generally a feature of introns. The usual dicot exonic 25–27% T content increases to around 42% for T at the expense of C and G (Simpson and Filipwicz, 1996). For site four, the sequence is 52% T, while for site 5 it is 49% T. The final possible site considered has only 5 amino acids before a stop codon is reached. Thus no reliable evidence was found for a possible 5' splice site to the proposed intron.

The other motif identified by the GENSCAN programme was a potential TCS near the beginning of A41. This reads CTCCATCA and corresponds well with the CTCATCA plant consensus TCS. Furthermore, there is a possible translation start site 67–75bp downstream reading AAA**ATG**GTA (highlighted in Figure 7.16), similar to the dicot consensus of AAA**ATG**GCTA (Joshi *et al.*, 1997). However, there are two problems here. Firstly the stop codon is reached without any indication of a 5' splice site and secondly there is a double microsatellite $TC_{(23)}TA_{(16)}$ 9bp upstream of the putative TCS. This would interrupt the usual positioning of the CAAT and TATA boxes, found around –70bp and –30bp relative to the TCS respectively (see Hughes, 1996), unless part of the microsatellite is acting as the TATA sequence. Microsatellites have been shown to act as enhancers within the promoter region, but they appear to be located further upstream than this one would be.

There are two possible conclusions to these observations. The first is that A41 contains a particularly large intron and the extent of the sequencing did not reach the 5' splice site. Intron sizes for other cassava genes vary in size, from 79–244bp for the 12

introns found in the CBG3 β-glucosidase gene (Liddle *et al.*, 1997) to 711bp for the single intron found in the gene for phenylalanine ammonia lyase (PAL; Y. Han, unpublished observations). However, the intron here would need to be 1667bp, which seems unlikely. Alternatively, A41 may in fact be a single exon gene and not the terminal exon. Investigating this possibility, the first methionine residue of the open reading frame is found at 1683–1685bp of the complete genomic sequence. Overall, plant genes tend to be A-T rich at the 5' ends. The AUG at the start of the mRNA, encoding the initial methionine, is frequently followed by G at position 4 and C at position 5. This results in a methionine-alanine start to many plant peptides (Heidecker and Messing, 1986) and this has been seen in cassava genes for PAL (Y. Han, unpublished observations), GBSS I and GBSS II (Munyikwa *et al.*, 1997). However, other cassava genes show different residues at the second position. The hydroxynitrile lyase gene (HNL; Hughes *et al.*, 1994) starts methionine-valine, while genes for UDP-glucosyltransferases (Hughes and Hughes, 1994) start methionine-glycine, although these do only represent single base changes; GTA and GGA respectively, compared with the GCA for alanine. The context of the proposed cassava starting methionine (TCT**ATG**CATA) does not appear to fit the dicot consensus (AAA**ATG**GCTA; Joshi *et al.*, 1997). However, further examination of the region upstream to this appears to indicate possible regulatory sequences expected within promoters.

7.4.2 Regulatory sequences for gA41

Two possible transcription start sites are indicated, both located within 120bp of the translation start. The most 5' motif of CTCATTT (1566–1672bp) is perhaps more likely, as it compares more favourably with the CTCATCA consensus than the second motif of CACCAACT (1606–1613bp).

The TATA box, which binds TFIID and subsequently allows the binding of other factors and RNA polymerase II for the transcription of eukaryotic mRNAs, is the short

sequence TATAAAT or similar, located between 16 and 54 nucleotides upstream of the transcription start site and present in virtually all expressed plant genes (Hughes, 1996). The *CBG3* gene has the sequence TATATAAA (Liddle *et al.*, 1997). A potential TATA box for A41, TATATA, is shown highlighted at 1552–1559bp. Either of the TCSs proposed is within this distance. Located 16bp still further 5' is the motif CCAAT, a possible CAAT box and similar to that of HNL which is CCAAAT. This is usually located around 74bp upstream of the TCS, which fits well with the second TCS, at 1606–1613bp. These also compare with the positions of the putative TATA and CAAT boxes in the cassava gene *MeHNL4* at –34bp and –51bp respectively (Hughes *et al.*, 1998). Assuming these elements to be correctly identified, there is a poly (A) stretch between them and the translation start. This may function as another control element, as proteins are known that bind the poly (A) tail from mRNA studies (Jackson and Standart, 1990).

The polyadenylation signal, AAUAAA, is found between 15 and 23 nucleotides upstream of the end of the mature mRNA and whilst it is extremely well conserved in animals, this is not so in plants where it deviates widely and may be found as multiple copies. Only one copy appears to be present within A41, though an earlier possible signal (AATAAT) is identified at 2504–2509bp inclusive.

Despite the computer predictions that A41 represents a terminal exon, detailed analysis of the entire sequence appears to indicate that it is in fact a single exon gene. This might also explain why the BLAST searches done with the cDNA sequence identified the same areas in the same proteins as when the whole genomic sequence was submitted against the databases. The SMART results indicate a helix-turn-helix motif and BLAST searches identified this region as being similar to those from a number of plant and animal transcription factors. Hence A41 appears to be the sequence for a cassava DNA-binding protein, possibly even a transcription factor, although further experiments would be needed to confirm its identity.

The most similar sequence, that from *Arabidopsis thaliana*, did not align further in the 5' direction, but perhaps this is not surprising, as transcription factors vary widely in sequence, depending on the target DNA sequence. Homology is generally within the actual DNA-binding region, with the first helix positioned across the major groove of the DNA, with the second, or recognition helix, lying partly in the major groove where it makes specific contacts with the bases (for review, see Latchman, 1990). Hence variation in this second helix directs sequence-specific binding.

CHAPTER 8: GENERAL DISCUSSION

The aim of this project was to isolate and characterize the promoter to a cassava gene expressed in the storage root and corresponding to a 25kDa protein, previously found to be up-regulated by nitrogen. However, the unreliability of the antibody to this protein meant that an alternative route was followed, although still with the intention of finding a useful promoter.

Several clones obtained from the cDNA library, constructed from storage root parenchyma tissue and screened by PCR, were of interest. The most interesting were a cassava homologue for the α-NAC sequence, the protein for which is involved in binding RNA as it emerges from the ribosome, and the sequence corresponding to an apparently novel cassava DNA binding protein. This latter was used to probe a cassava genomic library and the resulting clone sequenced for a total of 3.67kb to investigate the promoter region.

Based on comparison with known regulatory sequences, from cassava where available and those from plants in general, particularly dicots, where cassava-specific information was not available, it appears that the genomic clone is that of a single exon gene. The transcription site and translation start do vary considerably from the dicot consensus sequences but this is balanced against evidence of a microsatellite interrupting the promoter in one of the more critical stretches just upstream of the proposed transcription start. Additionally there is no reliable evidence of an initial exon with intronic 5' splice site. As a single copy gene, showing little similarity with any of the accessions from a number of databases of DNA or protein sequences, it is hard to confidently identify this sequence. However, it has both a coil and a helix-loop-helix motif, and in the latter region shows reasonable homology with a number of transcription factors. Hence it may represent the first identified transcription factor from

cassava. Since the RT-PCR result was obtained using leaf RNA, this would not be a root-specific protein. However, it is expressed in the storage root under normal conditions and in non-deteriorating tissue, and it would be interesting to identify the target gene(s). Once this was known it could then be determined whether or not this particular promoter would be useful in further genetic experiments.

As it is difficult to be completely certain that the promoter region has been reached for A41, the cassava α-NAC sequence could be used to provide a promoter. This is also expressed under normal conditions in the storage root and its function of binding nascent RNA would indicate expression during all developmental stages. It has the advantage over the unknown gene A41 that it has already been studied and it would therefore make the identification of the control sequences easier.

One of the aims of obtaining promoters is to allow the expression of inserted genes within transgenic plants and a number of cassava promoters are currently being investigated, corresponding to genes for the starch biosynthetic enzymes, cyanogenesis and post-harvest physiological deterioration (R. Visser and colleagues; M. Hughes and colleagues; J. Beeching and colleagues). These represent the major areas of interest in cassava and it may be unnecessary to investigate others. However, with the current public lack of support for genetic manipulation it may become more important to up-regulate the genes for naturally expressed products to achieve goals such as the manipulation of nutrient composition in the storage root. In this case, a promoter for a gene that is expressed naturally in all areas of the plant may still be useful. Alternatively, the protein itself may regulate a gene or genes of importance, such as those mentioned above, allowing indirect manipulation.

As shown by the results of the nutrient study, storage roots in cassava are not produced by the plant under all conditions. Unlike most plants that form tubers, the storage root is expendable in times of good nutrient supply and this may have implications for altering the storage root composition, as well as for the direct application of fertilizer. There appears to be a feedback mechanism operating once a

particular level of sulphur and nitrogen is reached. Whether this mechanism is triggered as nutrients enter the root or whether internal accumulation is detected is unknown. Perhaps the level of the sulphur-containing amino acids forms part of the signal. If this were the case, increasing the amount of protein in the storage root could be difficult. There is clearly a lot more research to be done to even start answering these questions.

REFERENCES

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J.D. (1994) *Molecular Biology of the Cell* (3rd edition). New York, USA: Garland Publishing, Inc.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 401–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, Z., Miller, W. & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402.

Amersham (1993) cDNA rapid cloning module – λMOSElox.

Amersham (1994a) cDNA synthesis module.

Amersham (1994b) cDNA rapid adaptor ligation module.

Amersham (1994c) λ-DNA in vitro packaging module.

Amersham (1994d) cDNA rapid cloning module – λgt11.

Averrre, C.W. (1967) Vascular streaking of stored cassava roots. In: Proceedings of the 1st International Symposium of Tropical Root Crops, Trinidad. pp 31–35.

Avise, J.C. (1994) *Molecular markers, natural history and evolution.* New York, USA; Chapman and Hall.

Balagopalan, C. & Sundar, P. (1997) Integrated technology for the treatment and management of cyanoglucosides in cassava starch and sago factory waste waters. *African Journal of Root and Tuber Crops* **2**, 69–73.

Beeching, J.R., Han, Y., Gómez-Vásquez, R., Day, R.C. & Cooper, R.M. (1998) Wound and defense responses in cassava related to post-harvest physiological deterioration. In: Phytochemical Signals and Plant-Microbe Interactions [Ed. Romeo *et al.*] pp 231–249. New York, USA; Plenum Press.

Beeching, J.R., Dodge, A.D., Moore, K.G. & Wenham, J.E. (1994) Physiological deterioration in cassava: an incomplete wound response? In: The Cassava

Biotechnology Network Proceedings of the 2nd International Scientific Meeting, Bogor, Indonesia 22–23 August 1994. pp 729–736.

Beezhold, D.H., Kostyal, D.A., Hickey, V.L., Noti, J.D. & Sussman, G.L. (1997) Characterization of latex allergen with homology to patatin. *Journal of Allergy and Clinical Immunology* **99**, 1088.

Bennett, R.N. & Wallsgrove, R.M. (1994) Tansley Review No. 72. Secondary metabolites in plant defence mechanisms. *New Phytologist* **127**, 617–633.

Benson, L. (1979) Plant classification (2nd edition). Massachussetts, USA, Toronto, Canada; DC Heath and Company.

Bertram, R.B. (1991) Cassava. In: Agricultural Biotechnology. Opportunities for International Development. [Ed. Persley, G.J.] Wallingford, UK; CAB International.

Boher, B. & Verdier, V. (1994) Cassava bacterial blight in Africa: the state of knowledge and implications for designing control strategies. *African Crop Science Journal* **2**, 505–509.

Bokanga, M. (1995) Biotechnology and Food Processing in Africa. *Food Technology* **49**, 86–90.

Bradshaw, H.D., Hollick, J.B., Parsons, T.J., Clarke, H.R.G. & Gordon, M.P. (1989) Systemically wound-responsive genes in poplar trees encode proteins similar to sweet potato sporamins and legume Kunitz trypsin inhibitors. *Plant Molecular Biology* **14**, 51–59.

Burge, C. & Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**, 78–94.

Caballero Vera, H., Villafuerte, J. & Poats, S.V. (1994) Transfer and adaptation of Colombian sour starch technology to Uatappy cassava producer processors in Manabi, Ecuador. *African Journal of Root and Tuber Crops* **2**, 253–257.

Chase, M.W., Soltis, D.E., Olmstead, R.G. *et al.* (1993) Phylogenetics of seed plants:

an analysis of nucleotide sequences from the plastid gene *rbc*L. *Annals of the Missouri Botanical Gardens* **80**, 581–586.

Chiwona-Karltun, L., Ngoma, J., Mahungu, N.M., Saka, J., Tylleskär, T. & Rosling, H. (1997) Reasons for use of 'bitter' and toxic cassava in Malawi. *African Journal of Root and Tuber Crops* **2**, 95–96.

Clark, J.M. (1988) Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic enzymes. Nucleic Acids Research **16**, 9677.

Clowes, A.E.E., Tatham, A.S., Beeching, J.R. & Shewry, P.R (1994) Characterization of cassava root proteins. *The Cassava Biotechnology Network; Proceedings of the Second International Scientific Meeting, Bogor, Indonesia* 716–728.

Cock, J.H. (1985) *Cassava: New potential for a neglected crop.* Boulder, CO, USA: Westfield Press.

Conlan, S., Griffiths, L-A., Turner, M., Fido, R., Tatham, A., Ainsworth, C. & Shewry, P. (1998) Characterisation of the yam tuber storage protein dioscorin. *Journal of Plant Physiology* **153**, 25–31.

Conlan, R.S., Griffiths, L-A., Napier, J.A., Shewry, P.R., Mantell, S. & Ainsworth, C. (1995) Isolation and characterisation of cDNA clones representing the genes encoding the major tuber storage protein (dioscorin) of yam (*Dioscorea cayenensis* Lam.) *Plant Molecular Biology* **28**, 369–380.

Conn, E.E. (1980) Cyanogenic compounds. *Annual Review of Plant Physiology* **31**, 433–451.

Cooper, R.M., Day, R., Burgin, R., Rodgers-Gray, B., Flood, J., Mepsted, R., Henshaw, G.G., Taylor, N., Gomez-Vasquez, R. & Beeching, J.R. (1997) Screening for and basis of resistance to cassava bacterial blight (*Xanthomonas campestris* pv. *manihotis*). *African Journal of Root and Tuber Crops* **2**, 54–55.

Dahniya, M.T. (1994) An overview of cassava in Africa. *African Crop Science Journal* **2**, 337–343.

Darwin, C.R. (1859) On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (Xth edition). Cambridge, UK; Cambridge University Press.

De Rijk, P., Robbrecht, E., de Hoog, S., Caers, A., Van de Peer, Y., De Wachter, R. (1999) Database on the structure of large subunit ribosomal RNA. *Nucleic Acids Research* **27**, 174–178.

Dellaporta, S.L., Wood, J., Hicks, J.B. (1983) A plant DNA mini preparation: version II. *Plant Molecular Biology Reports* **1**, 19–21.

Deng, D., Otim-Nape, W.G., Sangare, A., Ogwal, S., Beachy, R.N. & Fauquet, C.M. (1997) Presence of a new virus closely related to East African cassava mosaic geminivirus, associated with cassava mosaic outbreak in Uganda. *African Journal of Root and Tuber Crops* **2**, 23–28.

Du, L.C., Bokanga, M., Moller, B.L. & Halkier, B.A. (1995) The biosynthesis of cyanogenic glucosides in roots of cassava. *Phytochemistry* **39**, 323–326.

Dufour, D.L. (1988) Cyanide content of cassava (*Manihot esculenta*, Euphorbiaceae) cultivars used by Tukanoan Indians in Northwest Amazonia. *Economics of Botany* **42**, 255–266.

Emes, M.J. & Tobin, A.K. (1993) Physical and Molecular Structure of Plastids. [Eds Jeon, K.W. & Jarvik, J.] In: International Review of Cytology / A survey of Cell Biology **145**,174–185.

Escalada, R.G. & Ratilla, B.C. (1998) Effects of *Leucaena* biomass application in conjunction with fertilizers on cassava and taro yields in the Philippines. *Agroforestry Systems* **41**, 251–266.

FAO (Food and Agricultural Organization) (2000) FAOSTAT database http://apps.fao.org/

Fargette, D., Fauquet, C. & Thresh, J.M. (1994) Analysis and modelling of the temporal spread of African cassava mosaic virus and implications for disease control.

African Crop Science Journal **2**, 449–458.

Felstenstein, J. (1993) PHYLIP (Phylogeny Inferencing Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, USA.

Fowler, C. & Mooney, P. (1990) The Threatened Gene. Food, politics and the loss of genetic diversity. Cambridge, UK; The Lutterworth Press.

Fox, G.E., Magrum, L.J., Balch, W.E., Wolfe, R.S. & Woese, C.R. (1977) Classification of methanogenic bacteria by 16s ribosomal RNA characterization. *Proceedings of the National Academy of Sciences USA* **74**, 4537–4541.

Fraser, D.C, Craigmile, S. & Russell, G.C. (1994) Accelerated screening of cDNA libraries using the polymerase chain reaction and southern blotting. *Genetic Analysis Techniques and Application* **11**, 87–89.

Fregene, M.A., Vargas, J., Ikea, J., Angel, F., Tohme, J., Asiedu, R.A., Akoroda, M.O. & Roca, W.M. (1994) Variability of chloroplast DNA and nuclear ribosomal DNA in cassava (*Manihot esculenta* Crantz) and its wild relatives. *Theoretical and Applied Genetics* **89**, 719–727.

Fregene, M.A., Angel, F., Gómez, R., Rodriguez, F., Chavarriaga, P., Roca, W., Tohme, J. & Bonierbale, M. (1997) A molecular genetic map of cassava (*Manihot esculenta* Crantz). *Theoretical and Applied Genetics* **95**, 431–441.

Frohman, M.A., Dush, M.K. & Martin, G.R. (1988) Rapid amplification of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. *Proceedings of the National Academy of Sciences USA* **85**, 8988–9002.

Fulton, M.C., Bell, L.C. & Asher, C.J. (1996) Mineral nutrition of cassava (*Manihot esculenta* Crantz) grown in replaced soil after bauxite mining at Weipa, Queensland. *Australian Journal of Experimental Agriculture* **36**, 905–912.

Ghisolfi-Nieto, L., Joseph, G., Puvion-Dutilleul, F., Amalric, F. & Bouvet, P. (1996)

Nucleolin is a sequence-specific RNA-binding protein: characterization of targets on pre-ribosomal RNA. *Journal of Molecular Biology* **260**, 34–53.

Hahn, S.K. (1989) An overview of African traditional cassava processing and utilisation. *Outlook on Agriculture* **18**, 110–118.

Hahn, S.K., Terry, E.R. & Leuschner, K. (1980) Cassava breeding for resistance to cassava mosaic disease. *Euphytica* **29**, 673–683.

Hamby, R.K. and Zimmer, E.A. (1992) Ribosomal RNA as a phylogenetic tool in plant systematics. *In*: Molecular Systematics of Plants, pp 50–91. [Soltis, P.S., Soltis, D.E. & Doyle, J.J., eds] New York, USA; Chapman and Hall.

Harvey, P.J. & Boulter, D. (1983) Isolation and characterisation of the storage proteins of yam tubers. *Phytochemistry* **22**, 1687–1693.

Hattori, T., Nakagawa, T., Maeshima, M., Nakamura, K. & Asahi, T. (1985) Molecular cloning and nucleotide sequence of cDNA for sporamin, the major soluble protein of sweet potato tuberous roots. *Plant Molecular Biology* **5**, 313–320.

Hattori, T., Matsuoka, K. & Nakamura, K. (1988) Subcellular localization of the sweet potato tuberous root storage protein. *Agricultural Biological Chemistry* **52**, 1057–1059.

Hattori, T., Yoshida, N. & Nakamura, K. (1989) Structural relationship among the members of a multigene family coding for the sweet potato tuberous roots storage protein. *Plant Molecular Biology* **13**, 563–572.

Heidecker, G. & Messing, J. (1986) Structural analysis of plant genes. *Annual Review of Plant Physiology* **37**, 439–466.

Henry, G., Titapiwatanakun, B., Bottema, T. & Damardjati, D.S. (1994) Asian cassava market dynamics: opportunities for biotechnology. In: The Cassava Biotechnology Network Proceedings of the 2nd International Scientific Meeting, Bogor, Indonesia 22–23 August 1994. pp 581–608.

Herren, H.R. (1994) Cassava pest and disease management: an overview. *African*

Crop Science Journal **2**, 345–353.

Hughes, M.A. (1993) Molecular Genetics of Plant Cyanogenic β-Glucosidases. In: β-Glucosidases: Biochemistry and Molecular Biology pp 153–169.

Hughes, M.A. (1996) *Plant Molecular Genetics.* Harlow, UK; Longman.

Hughes, J. & Hughes, M.A. (1994) Multiple secondary plant product UDP-glucose glucosyltransferase genes expressed in cassava (*Manihot esculenta*, Crantz) cotyledons. *DNA Sequence* **5**, 41–49.

Hughes, M.A., Brown, K., Pancoro, A., Murray, B.S., Oxtoby, E. & Hughes, J. (1992) A molecular and biochemical analysis of the structure of the cyanogenic β-glucosidase, linamarase, from cassava (*Manihot esculenta*, Crantz). *Archives of Biochemistry and Biophysics* **295**, 273–279.

Hughes, J., Carvalho, F.J.P.C. & Hughes, M.A. (1994) Purification, characterisation and cloning of α-hydroxynitrile lyase from cassava (*Manihot esculenta*, Crantz). *Archives of Biochemistry and Biophysics* **311**, 496–502.

Hughes, J., Keresztessy, Z., Brown, K., Suhandono, S. & Hughes, M.A. (1998) Genomic organization and structure of α-hydroxynitrile lyase in cassava (*Manihot esculenta* Crantz). *Archives of Biochemistry and Biophysics* **356**, 107–116.

Hunt L.A., Wholey, D.W. & Cock, J.H. (1977) Growth physiology of cassava. *Field Crop Abstracts* **30**, 77–91.

IITA (1990) Root, Tuber and Plantain Improvement. In: International Institute of Tropical Agriculture Annual Report 1989/90 pp 41–50.

Ishiguro, S. & Nakamura, K. (1992) The nuclear factor SP8BF binds to the 5'-upstream regions of the three different genes coding for major proteins of sweet potato tuberous roots. *Plant Molecular Biology* **18**, 97–108.

Jackson, R.J. & Standart, N. (1990) Do the poly (A) tail and 3' untranslated region control mRNA translation? *Cell* **62**, 15–24.

Joshi, C.P., Zhou, H., Huang, X. & Chiang, V.L. (1997) Context sequences of translation initiation codon in plants. *Plant Molecular Biology* **35**, 993–1001.

Kapinga, R., Mlingi, N. & Rosling, H. (1997) Reasons for use of bitter cassava in southern Tanzania. *African Journal of Root and Tuber Crops* **2**, 81–84.

Koch, B., Nielsen, V.S., Halkier, B.A., Olsen, C.E., Møller, B.L. (1992) The biosynthesis of cyanogenic glycosides in seedlings of cassava (*Manihot esculenta*, Crantz). *Archives of Biochemistry and Biophysics* **292**, 141–150.

Kostyal, D.A., Hickey, V.L., Noti, V.D., Sussman, G.L. & Beezhold, D.H. (1998) Cloning and characterization of a latex allergen (Hev b 7): homology to patatin, a plant PLA (2). *Clinical and Experimental Immunology* **112**, 355–362.

Larson, R.A. (1995) Plant defenses against oxidative stress. *Archives of Biochemistry and Physiology* **29**, 175–186.

Latchman, D.S. (1990) Gene regulation: a eukaryotic perspective. London, UK; Unwin Hyman Ltd.

Learned, R.M., Learned, T.K., Haltiner, M.M & Tijan, R.T. (1986) Human rRNA transcription is modulated by the coordinate binding of two factors to an upstream control element. *Cell* **45**, 847–857.

Lee, S.B. & Taylor, J.W. (1992) Phylogeny of five fungus-like protoctistan *Phytophthora* species, inferred from the internal transcribed spacers of ribosomal DNA. *Molecular Biology of Evolution* **9**, 636–653.

Li, H.Q., Sautter, C., Potrykus, I. & Puonti-Kaerlas, J. (1996) *Nature Biotechnology* **14**, 736–740.

Liddle, S., Hughes, J. & Hughes, M.A. (1997) Analysis of a cassava root-specific β-glycosidase promoter. *African Journal of Root and Tuber Crops* **2**, 158–162.

Linn, S. & Arber, W. (1968) Host specificity of DNA produced by *Escherichia coli*. X. *In vitro* restriction of phage fd replicative form. *Proceedings of the National Academy of Sciences USA* **59**, 1300–1306.

Loeb, L.A. & Kunkel, T.A. (1982) Fidelity of DNA synthesis. *Annual Review of Biochemistry* **51**, 429–457.

Mabberley, D.J. (1997) The Plant-Book. A portable dictionary of the vascular plants. (2nd edition) Cambridge, UK; Cambridge University Press.

Maeshima, M., Sasaki, T. & Asahi, T. (1985) Characterization of major proteins in sweet potato tuberous roots. *Phytochemistry* **24**, 1899–1902.

Maxam, A.M. & Gilbert, W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences USA* **74**, 560–564.

McMahon, J.M., White, W.L.B. & Sayre, R.T. (1995) Cyanogenesis in cassava (*Manihot esculenta*, Crantz). *Journal of Experimental Botany* **46**, 731–741.

Meselson, M. & Yuan, R. (1968) DNA restriction enzyme from *E. coli*. *Nature* **217**, 1110–1114.

Mignery, G.A., Pikaard, C.S. & Park, W.D. (1984) Molecular characterization of patatin multigene family of potato. *Gene* **62**, 27–44.

Monte-Neshich, D.C., Rocha, T.L., Guimarães, R.L., Santana, E.F., Loureiro, M.E., Valle, M. & Grossi de Sá, F. (1995) Characterization and spatial localization of the major globulin families of taro (*Colocasia esculenta* L. Schott) tubers. *Plant Science* **112**, 149–159.

Munyikwa, T.R.I., Chipangura, B., Salehuzzaman, S.N.I.M., Jacobsen, E. & Visser, R.G.F. (1994) Cloning and characterization of cassava genes involved in starch biosynthesis. In: The Cassava Biotechnology Network Proceedings of the 2nd International Scientific Meeting, Bogor, Indonesia 22–23 August 1994. pp 639–645. Cali, Colombia; CIAT.

Munyikwa, T.R.I., Jacobsen, E. & Visser, R.G.F. (1997) Developmental regulation of cassava granule bound starch synthase II. *African Journal of Root and Tuber Crops* **2**, 116–120.

Neuenschwander, P. (1994) Control of the cassava mealybug in Africa: lessons from a

biological control project. *African Crop Science Journal* **2**, 369–383.

Ohto, M., Nakamura-Kito, K. & Nakamura, K. (1992) Induction of expression of genes coding for sporamin and β-amylase by polygalcturonic acid in leaf-petiole cuttings of sweet potato. *Plant Physiology* **99**, 422–427.

Olasantan, F.O., Ezumah, H.C. & Lucas, E.O. (1997) Response of cassava and maize to fertilizer application, and a comparison of the factors affecting their growth during intercropping. *Nutrient Cycling in Agroecosystems* **46**, 215–223.

Olsen, K.M. & Schaal, B.A. (1999) Evidence on the origin of cassava: Phylogeography of *Manihot esculenta*. *Proceedings of the National Academy of Sciences USA* **96**, 5586–5591.

Otim-Nape, G.W., Bua, A. & Baguma, Y. (1994) Accelerating the transfer of improved production technologies: controlling African cassava mosaic virus disease epidemics in Uganda. *African Crop Science Journal* **2**, 479–495.

Page, R.D.M. & Holmes, E.C. (1998) *Molecular evolution. A phylogenetic approach.* Oxford, UK; Blackwell Science Ltd.

Park, W.D. (1983) Tuber proteins of potato – a new and surprising molecular system. *Plant Molecular Biology Reported* **1**, 61–66.

Pawlowski, K., Kunze, R., de Vries, S. & Bisseling, T. (1994) Isolation of total, poly (A) and polysomal RNA from plant tissues. In: Gelvin, S.B. & Schilperoort, R.A. (eds), Plant Molecular Biology Manual. Amsterdam, The Netherlands; Kluwer Academic Publishers.

Pearson, W.R. & Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences USA* **85**, 2444–2448.

Persley, G.J. & Doyle, J.J. (1999) *Biotechnology for developing-country agriculture: problems and opportunities.* 2020 Overview. Washington DC, USA; IFPRI.

Peterson, S., Legue, F., Tylleskär, T., Kpizingui, E. & Rosling, H. (1995) Improved cassava processing can help reduce iodine deficiency disorders in the Central African Republic. *Nutrition Research* **15**, 803–812.

Pikaard, C.S., Mignery, G.A., Ma, D.P., Stark, V.J. & Park, W.D. (1986) Sequence of two apparent pseudogenes of the major potato tuber protein patatin. *Nucleic Acids Research* **12**, 5564–5566.

Pikaard, C.S., Brusca, J.S., Hannapel, D.J. & Park, W.D. (1987) The two classes of genes for the major tuber protein, patatin, are differentially expressed in tubers and roots. *Nucleic Acids Research* **15**, 1979–1994.

Pinstrup-Anderson, P., Pandya-Lorch, R. & Rosegrant, M.W. (1999) *World food prospects: critical issues for the early twenty-first century*. 2020 Food Policy Report. Washington DC, USA; IFPRI.

Plumbley, R.A., Hughes, P.A. & Marriott, J. (1981) Studies on peroxidases and vascular discoloration in cassava root tissue. *Journal of the Science of Food and Agriculture* **32**, 723–731.

Prat, S., Frommer, W.B., Höfgen, R., Keil, M., Koßmann, J., Köster-Töpfer, M., Liu, X-J., Müller, B., Peña-Cortés, H., Rocha-Sosa, M., Sánchez-Serrano, J.J., Sonnewald, U. & Willmitzer, L. (1990) Gene expression during tuber development in potato plants. *FEBS Letters* **268**, 334–338.

Puonti-Kaerlas, J. (1998) Cassava Biotechnology. *Biotechnology and Genetic Engineering Reviews* **15**, 329–364.

Racusen, D. (1984) Lipid acyl hydrolase of patatin. *Canadian Journal of Botany* **62**, 1640–1644.

Raemakers, C.J.J.M., Schavemaker, C.M., Jacobsen, E. & Visser, R.G.F. (1993) Improvements of cyclic somatic embryogenesis of cassava (*Manihot esculenta* Crantz). *Plant Cell Reports* **12**, 226–229.

Reimann, B., Bradsher, J., Franke, J., Hartmann, E., Wiedmann, M., Prehn, S. & Wiedmann, B. (1999) Initial charaterization of the nascent polypeptide-associated complex in yeast. *Yeast* **15**, 397–407.

Restrepo, S., Verdier, V., Mosquera, G., Gerstl, A., Laberry, R., Valle, T. & Alvarez, E. (1997) Cassava bacterial blight in South America: pathogenic and genetic characterization of the causal agent and its application to screening methods. *African Journal of Root and Tuber Crops* **2**, 61–63.

Rickard, J.E. & Coursey, D.G. (1981) Cassava storage. Part 1: Storage of fresh cassava roots. *Tropical Science* **23**, 1–32.

Rickard, J.E. & Gahan, P.B. (1983) The development of occlusions in cassava (*Manihot esculenta*, Crantz) root xylem vessels. *Annals of Botany* **52**, 811–821.

Rogers, D.J. & Fleming, H.S. (1973) A monograph of *Manihot esculenta* with an explanation of the taximetric methods used. *Economic Botany* **27**, 1–113.

Rosahl, S., Schell., J & Willmitzer, L. (1987) Expression of a tuber-specific storage protein in transgenic tobacco plants: demonstration of an esterase activity. *EMBO Journal* **6**, 1155–1159.

Rothnie, H.M. (1996) Plant mRNA 3'-end formation. *Plant Molecular Biology* **32**, 43–61.

Ryan, C.A. (1981) *Proteinase inhibitors*. pp 351–370 [Eds Stumpf, P.K. & Conn, E.E.] In: The Biochemistry of Plants, Vol 6. New York; Academic Press.

Salehuzzaman, S.N.I.M., Jacobsen, E. & Visser, R.G.F. (1992) Cloning, partial sequencing and expression of a cDNA coding for branching enzyme in cassava. *Plant Molecular Biology* **20**, 809–819.

Salehuzzaman, S.N.I.M., Jacobsen, E. & Visser, R.G.F. (1993) Isolation and characterisation of a cDNA encoding granule-bound starch synthase in cassava (*Manihot esculenta*, Crantz) and its antisense expression in potato. *Plant Molecular Biology* **23**, 947–962.

Sanger, F., Nicklen, S. & Coulson, A.R. (1977) DNA sequencing with chain terminating

inhibitors. *Proceedings of the National Academy of Sciences USA* **74**, 5463–5467.

Schöpke, C., Taylor, N., Cárcamo, R., Konan, N.K., Marmey, P., Henshaw, G.G., Beachy, R. & Fauquet, C. (1996) Regeneration of transgenic cassava plants (*Manihot esculenta* Crantz) from microbombarded embryogenic suspension cultures. *Nature Biotechnology* **14**, 731–735.

Schultz, J., Milpetz, F., Bork, P. & Ponting, C.P. (1998) SMART, a simple modular architecture research tool: identification of signalling domains. *Proceedings of the National Academy of Sciences USA* **95**, 5857–5864.

Senda, K., Yoshioka, H., Doke, N. & Kawakita, K. (1996) A cytosolic phosholipase A(2) from potato tissue appears to be patatin. *Plant and Cell Physiology* **37**, 347–353.

Shewry, P.R. (1995) Plant storage proteins. *Biological Reviews of the Cambridge Philosophical Society* **70**, 375–426.

Shewry, P.R., Clowes, A., Tatham, A.S. & Beeching, J.R. (1993) pp 251–254. Opportunities for manipulating the amount and composition of the proteins of cassava tuberous roots. In: *Proceedings of the 1st International Scientific Meeting of the Cassava Biotechnology Network, Cartagena de Indias, Colombia, September 1992.* CIAT, Cali, Colombia.

Shi, X., Parthun, M.R. & Jaehning, J.A. (1995) The yeast EGD2 gene encodes a homologue of the alpha NAC subunit of the human nascent-polypeptide-associated complex. *Gene* **165**, 199–202.

Simpson, G.G. & Filipowicz, W. (1996) Splicing of precursors to mRNA in higher plants: mechanism, regulation and sub-nuclear organisation of the spliceosomal machinery. *Plant Molecular Biology* **32**, 1–41.

Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology* **98**, 503–517.

Souza, P.A.S., Gomes, E. & Campos, F.A.P. (1998) Tissue distribution and deposition

pattern of a cellulosic parenchyma-specific protein from cassava roots. *Brazilian Archives of Biology and Technology* **41**, 1–9.

Ssemakula, G.N., Baguma, Y.K., Otim-Nape, G.W., Bua, A. & Ogwal, S. (1997) Breeding for resistance to mosaic disease in Uganda. *African Journal of Root and Tuber Crops* **2**, 36–42.

Stamp, J.A. & Henshaw, G.G. (1982) Somatic embryogenesis in cassava. *Zeitschrift für Pflanzenphysiologie* **105**, 183–187.

Stamp, J.A. & Henshaw, G.G. (1987) Secondary somatic embryogenesis and plant regeneration in cassava. *Plant Cell Tissue and Organ Culture* **10**, 227–233.

Sturz, A.V., Christie, B.R. & Matheson, B.G. (1998) Associations of bacterial endophyte populations from red clover and potato crops with potential for beneficial allelopathy. *Canadian Journal of Microbiology* **44**, 162–167.

Suh, S-G., Stiekema, W.J. & Hannapel, D.J. (1991) Proteinase-inhibitor activity and wound-inducible gene expression of the 22-kDa potato tuber proteins. *Planta* **184**, 423–430.

Taiz, L. & Zeiger, E. (1991) *Plant Physiology*. Redwood City, CA, USA: The Benjamin/Cummings Publishing Company, Inc.

Taylor, N.J., Edwards, M., Kiernan, R.J., Davey, C.D.M., Blakesley, D. & Henshaw, G.G. (1996) Development of friable embryogenic callus and embryogenic suspension culture systems in cassava (*Manihot esculenta* Crantz). *Nature Biotechnology* **14**, 726–730.

Thresh, J.M., Fargette, D. & Otim-Nape, G.W. (1994) The viruses and virus diseases of cassava in Africa. *African Crop Science Journal* **2**, 459–478.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.

Tor-Agbidye, J., Palmer, V.S., Lasarev, M.R., Craig, A.M., Blythe, L.L., Sabri, M.I. & Spencer, P.S. (1999) Bioactivation of cyanide to cyanate in sulfur amino acid deficiency: relevance to neurological disease in humans subsisting on cassava. *Toxicological Science* **50**, 228–235.

Tylleskär, T., Banea, M., Bikangi, N., Cooke, R.D., Poulter, N.H. & Rosling, H. (1992) Cassava cyanogens and konzo, an upper motoneuron disease found in Africa. *Lancet* **339**, 208–211.

Van de Peer, Y. & De Wachter, R. (1993) TREECON: a software package for the construction and drawing of evolutionary trees. *Computational Applications in Bioscience* **9**, 177–182.

Van de Peer, Y., Robbrecht, E., de Hoog, S., Caers, A., De Rijk, P. & De Wachter, R. (1999) Database on the structure of small subunit ribosomal RNA. *Nucleic Acids Research* **27**, 179–183.

van de Wal, M., D'Hulst, C., Vincken, J.P., Buleon, A., Visser, R. & Ball S. (1998) Amylose is synthesized in vitro by extension of and cleavage from amylopectin. *Journal of Biological Chemistry* **273**, 22232–22240.

van den Bos, R. (1997) The cassava biotechnology network: an integral part of the special programme's approach to reduce small farmers constraints. *African Journal of Root and Tuber Crops* **2**, 4–6.

Verdier, V., Restrepo, S., Boher, B., Nicole, M., Geiger, J.P., Alvarez, E. & Bonierbale, M. (1997) Cassava bacterial blight: recent achievements in understanding the disease. *African Journal of Root and Tuber Crops* **2**, 64–68.

Wheatley, C.C. & Schwabe, W.W. (1985) Scopoletin involvement in post-harvest deterioration of cassava root. *Journal of Experimental Botany* **36**, 783–791.

Wiedmann, B., Sakai, H., Davis, T.A. & Wiedmann, M. (1994) A protein complex required for signal-sequence-specific sorting and translocation. *Nature* **370**, 434–440.

Woese, C.R., Kandler, O. & Wheelis, M.L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eukarya. *Proceedings of the National Academy of Sciences USA* **87**, 4576–4579.

Yaninek, Y.S. (1994) Cassava green mite intervention technologies. *African Crop Science Journal* **2**, 361–367.

Yeh, K-W., Chen, J-C., Lin, M-I., Chen, Y-M & Lin, C-Y. (1997) Functional activity of sporamin from sweet potato (*Ipomoea batatas* Lam.): a tuber storage protein with trypsin inhibitory activity. *Plant Molecular Biology* **33**, 565–570.

Yeoh, H.H. & Chew, M.Y. (1977) Electrophoretic patterns of the soluble protein from cassava (*Manihot esculenta* Crantz) leaf. In: *Proceedings of the Malaysian Biochemical Society*, 4th Conference.

Zhao, F.J., Withers, P.J.A., Evans, E.J., Monaghan, J., Salmon, S.E., Shewry, P.R. & McGrath, S.P. (1997) Sulphur nutrition: An important factor for the quality of wheat and rapeseed. (Reprinted from *Plant nutrition for sustainable food production and environment*, 1997.) *Soil Science and Plant Nutrition* **43**, 1137–1142.

Zhou, X., Liu, Y., Calvert, L., Munoz, C., Otim-Nape, G.W., Robinson, D.J. & Harrison, B.D. (1997) Evidence that DNA-A of a geminivirus associated with severe cassava mosaic disease in Uganda has arisen by interspecific recombination. *Journal of General Virology* **78**, 2101–11.

Appendix 1 Recipes for commonly used buffers and media

Buffers:

50× TAE stock:

Tris	242g
Glacial acetic acid	57.1ml
0.5M stock EDTA, pH8	100ml

Dissolve in distilled water to 1l
Autoclave 20min at 15psi

10× TBE stock:

Tris	108g
Boric Acid	55g
0.5M stock EDTA, pH8	2ml

Dissolve in distilled water to 1l
Adjust pH to 8 using HCl
Autoclave 20min at 15psi

10× TE buffer:

Tris	1.21g
EDTA	0.37g

Dissolve in distilled water to 100ml
Adjust pH to 7.5
Autoclave 20min at 15psi

Media:

Luria-Bertani (LB) medium:

Bacto-tryptone	10g
Bacto-yeast extract	5g
NaCl	10g

Dissolve in distilled water to 1l
Adjust pH to 7 using NaOH
Autoclave 20min at 15psi

NZCYM medium:

Bacto-yeast extract	5g
NaCl	5g
Casein hydrolysate	11g
$MgSO_4$	2g

Dissolve in distilled water to 1l
Adjust pH to 7 using NaOH
Autoclave 20min at 15psi

Appendix 2 Cassava cultivars used in this project

Cultivar	Origin	Female parent/common name
CMC 40	Colombia	n/a
MBra 337	Brazilian landrace	n/a
MCol 22	Colombian landrace	Uvita
MCol 1684	Colombian landrace	Charay
MVen 77	Venezuelan landrace	Venezuela
NGA2	Nigeria	TMS 30572

Appendix 3 Tables showing statistical values used to analyse the wet weights of roots produced by cassava plants grown under different nutrient regimes (Chapter 3)

A: Table of means of the roots produced by the four treatments, together with the standard deviation

Treatment	Mean	Standard deviation
LN–S	390.96	128.50
LN+S	529.80	290.80
HN–S	537.68	129.15
HN+S	152.28	133.13

B: 2-way ANOVA values produced by Minitab

Source	DF	SS	MS	F	P
N	1	79899	79899	2.36	0.140
S	1	91194	91194	2.69	0.116
N*S	1	412239	412239	12.17	0.002
Error	20	677395	33870		
Total	23	1260727			

DF: Degrees of freedom; SS: sum of squares; MS: mean square; F: F-value; P: probability

Appendix 4 Tree generated for the seven full length 26S ribosomal DNA sequences available, with the algal sequence *C. ellipsoidea* used as the outgroup to root the tree. Analysis was carried out as described in Chapter 6 for the partial sequence analysis, with the outfile from DNAPARS used in the CONSENSE programme, with 100 bootstraps and the input order randomized. Numbers at the forks indicate the number of times the group consisting of the species which are to the right of that fork occurred among the trees, out of 100 trees.