

47kD AND 31kD COCOA BEAN PROTEINS AND
A 67kD PRECURSOR THEREFOR

5 This invention relates to proteins and nucleic acids derived from or otherwise related to cocoa.

10 The beans of the cocoa plant (*Theobroma cacao*) are the raw material for cocoa, chocolate and natural cocoa and chocolate flavouring. As described by Rohan ("Processing of Raw Cocoa for the Market", FAO/UN (1963)), raw cocoa beans are extracted from the harvested cocoa pod, from which the placenta is normally removed, the beans are then "fermented" for a period of days, during which the beans are killed and a purple pigment is released from the cotyledons. During fermentation "unknown" compounds are formed which on roasting give rise to characteristic cocoa flavour. Rohan suggests that polyphenols and theobromine are implicated in the flavour precursor formation. After fermentation, the beans are dried, during which time the characteristic brown pigment forms, and they are then stored and shipped.

20 Biehl *et al*, 1982 investigated proteolysis during anaerobic cocoa seed incubation and identified 26kD and 44kD proteins which accumulated during seed ripening and degraded during germination. Biehl asserted that there were storage proteins and suggested that they may give rise to flavour-specific peptides.

25 Fritz *et al*, 1985 identified polypeptides of 20kD and 28kD appearing in the cytoplasmic fraction of cocoa seed extracts at about 100 days after pollination. It appears that the 20kD protein is thought to have glyceryl acyltransferase activity.

30 In spite of the uncertainties in the art, as summarised above, proteins apparently responsible for flavour production in cocoa beans have now been identified. Further, it has been discovered that, in spite of Fritz's caution that "cocoa seed

mRNA levels are notably low compared to other plants" (*loc. cit.*), it is possible to apply the techniques of recombinant DNA techniques to the production of such proteins.

5 According to a first aspect of the invention, there is provided a protein encoded by mRNA extractable from beans of cocoa (*Theobroma cacao*), the protein having a molecular weight of 67 kD, as assessed by SDS-PAGE, and at least part of the amino acid sequence shown in Figure 3 (including at least some of the unboxed residues in the sequence marked *Tca* in Figure 4), or a fragment of
10 such a protein having at least 20 amino acids.

The 67kD protein appears to be a primary translation product of interest in proteins involved in flavour production in cocoa. The 67kD protein may be processed in vivo to form 47kD and 31kD polypeptides.

15 According to a second aspect of the invention, there is provided a protein extractable from beans of cocoa (*Theobroma cacao*), the protein having a molecular weight of 47kD, as assessed by SDS-PAGE, and part of the amino acid sequence shown in Figure 3 (including at least some of the unboxed
20 residues in the sequence marked *Tca* in Figure 4), or a fragment of such a protein having at least 20 amino acids.

25 According to a third aspect of the invention, there is provided a protein extractable from beans of cocoa (*Theobroma cacao*), the protein having a molecular weight of 31kD, as assessed by SDS-PAGE, and part of the amino acid sequence shown in Figure 3 (including at least some of the unboxed residues in the sequence marked *Tca* in Figure 4), or a fragment of such a protein having at least 20 amino acids.

30 The term "fragment" as used herein and as applied to proteins or peptides indicates a sufficient number of amino acid residues are present for the fragment

to be useful. At least 20 amino acids are present in a fragment. Useful fragments include those which are the same as or similar or equivalent to those naturally produced during the fermentation phase of cocoa bean processing. It is believed that such fragments take part in Maillard reactions during roasting, to form at least some of the essential flavour components of cocoa.

Proteins in accordance with the invention may be synthetic; they may be chemically synthesised or, preferably, produced by recombinant DNA techniques. Proteins produced by such techniques can therefore be termed "recombinant proteins". Recombinant proteins may be glycosylated or non-glycosylated; non-glycosylated proteins will result from prokaryotic expression systems.

Theobroma cacao has two primary subspecies, *Th. cacao cacao* and *Th. cacao sphaerocarpum*. While proteins in accordance with the invention may be derived from these subspecies, the invention is not limited solely to these subspecies. For example, many cocoa varieties are hybrids between different species; an example of such a hybrid is the trinitario variety.

The invention also relates to nucleic acid, particularly DNA, coding for the proteins referred to above (whether the primary translation products, the processed proteins or fragments). The invention therefore also provides, in a further aspect, recombinant or isolated nucleic acid which comprises:

- (a) a nucleotide sequence as shown in Figure 2, encoding a 67, 47 or 31kD protein (as measured by SDS-PAGE); or
- (b) a nucleotide sequence which hybridises to (a); or
- (c) a fragment of (a) or (b) of at least 60 nucleotides.

Included in the invention is nucleic acid which is degenerate for the wild type protein and which codes for conservative or other non-deleterious mutants. Nucleic acid which hybridises to the wild type material is also included.

Frequently, nucleic acid in accordance with the invention will be incorporated into a vector (whether an expression vector or otherwise) such as a plasmid. Suitable expression vectors will contain an appropriate promoter, depending on the intended expression host. For yeast, an appropriate promoter is the yeast pyruvate kinase (PK) promoter; for bacteria an appropriate promoter is a strong lambda promoter.

Expression may be secreted or non-secreted. Secreted expression is preferred, particularly in eukaryotic expression systems; an appropriate signal sequence may be present for this purpose. Signal sequences derived from the expression host (such as that from the yeast alpha-factor in the case of yeast) may be more appropriate than native cocoa signal sequences.

The invention further relates to host cells comprising nucleic acid as described above. Genetic manipulation may for preference take place in prokaryotes. Expression will for preference take place in a food-approved host. The yeast strains *Saccharomyces cerevisiae* and *Hansenula polymorpha* are particularly preferred. *Escherichia coli* is also a suitable host cell for expression.

The invention also relates to processes for preparing nucleic acid and protein as described above by nucleic acid replication and expression, respectively.

cDNA in accordance with the invention may be useful not only for obtaining protein expression but also for Restriction Fragment Length Polymorphism (RFLP) studies. In such studies, detectably labelled cDNA (eg radiolabelled) is prepared. DNA of a cultivar under analysis is then prepared and digested with restriction enzymes. Southern blotting with the labelled cDNA may then enable genetic correlations to be made between cultivars. Phenotypic correlations may then be deduced.

5

The invention will now be illustrated by the following non-limiting examples. The examples refer to the accompanying drawings, in which:

5 Figure 1 shows a map of the coding region of the 67kD protein, together with the inter-relationship of plasmids pMS600, pMS700 and pMS800, from which sequence data were obtained;

10 Figure 2 shows the complete nucleotide sequence of cDNA coding for the 67kD protein and the deduced amino acid sequence;

Figure 3 shows the amino acid sequence referred to in Figure 2;

Figure 4 shows the relationship between the 67kD protein and seed storage proteins from other plants;

Figure 5 shows a map of plasmid pJLA502;

Figure 6 shows schematically the formation of plasmid pMS900;

Figure 7 shows two yeast expression vectors useful in the present invention; vector A is designed for internal expression and vector B is designed for secreted expression;

Figure 8a shows, in relation to vector A, part of the yeast pyruvate kinase gene showing the vector A cloning site, and the use of *Hin-Nco* linkers to splice in the heterologous gene;

Figure 8b shows, in relation to vector B, part of the yeast alpha-factor signal sequence showing the vector B cloning site, and the use of *Hin-Nco* linkers to create an in-phase fusion;

Figure 9a shows how plasmid pMS900 can be manipulated to produce plasmids pMS901, pMS903, pMS907, pMS908, pMS911, pMS912 and pMS914;

Figure 9b shows how plasmid pMS903 can be manipulated to produce plasmids pMS904, pMS905, pMS906, pMS909 and pMS916;

Figure 10 shows maps of plasmids pMS908, pMS914, pMS912, pMS906, pMS916 and pMS910;

Figure 11 shows the construction of a plasmid to express the 67kD protein by means of the AOX promoter on an integrated vector in *Hansenula polymorpha*; and

7

1 Figure 12 shows the construction of a plasmid to express the 67kD protein by
2 means of the AOX promoter in conjunction with the yeast α -factor secretory
3 signal on an integrated vector in *Hansenula polymorpha*.

4 5 EXAMPLES

6 7 Example 1

8 9 *Identification of the Major Seed Proteins*

10
11 It is not practicable to extract proteins directly from cocoa beans due to the high
12 fat and polyphenol contents, and proteins were, therefore, extracted from
13 acetone powders made as follows. Mature beans from cocoa of West African
14 origin (*Theobroma cacao amelonada*) were lyophilised and ground roughly in a
15 pestle and mortar. Lipids were extracted by Soxhlet extraction with diethyl
16 ether for two periods of four hours, the beans being dried and further ground
17 between extractions. Polyphenols and pigments were then removed by several
18 extractions with 80% acetone, 0.1% thioglycolic acid. After extraction the
19 resulting paste was dried under vacuum and ground to a fine powder.

20
21 Total proteins were solubilised by grinding the powder with extraction buffer
22 (0.05 M sodium phosphate, pH 7.2; 0.01 M 2-mercaptoethanol; 1% SDS) in a
23 hand-held homogeniser, at 5mg/ml. The suspension was heated at 95°C for 5
24 minutes, and centrifuged at 18 K for 20 minutes to remove insoluble material.
25 The resulting clear supernatant contained about 1 mg/ml total protein.
26 Electrophoresis of 25 μ l on an SDS-PAGE gel (Laemmli, 1970) gave three
27 major bands, two of which were at 47 kD and 31 kD, comprising over 60% of
28 the total proteins. The 47kD and 31kD proteins are presumed to be the
29 polypeptide subunits of major storage proteins.

Characteristics of the Storage Polypeptides

The solubility characteristics of the 47 kD and 31 kD polypeptides were roughly defined by one or two quick experiments. Dialysis of the polypeptide solution against SDS-free extraction buffer rendered the 47 kD and 31 kD polypeptides insoluble, as judged by their ability to pass through a 0.22 micron membrane. Fast Protein Liquid Chromatography (FPLC) analysis also showed that the 47 kD and 31 kD polypeptides were highly associated after extraction with McIlvaines buffer pH 6.8 (0.2 M disodium hydrogen phosphate titrated with 0.1 M citric acid). The 47 kD and 31 kD polypeptides are globulins on the basis on their solubility.

Purification of the 47 kD and 31 kD polypeptides

The 47 kD and 31 kD polypeptides were purified by two rounds of gel filtration on a SUPEROSE-12 column of the PHARMACIA Fast Protein Liquid Chromatography system (FPLC), or by electroelution of bands after preparative electrophoresis. (The words SUPEROSE and PHARMACIA are trade marks.) Concentrated protein extracts were made from 50 mg acetone powder per ml of extraction buffer, and 1-2 ml loaded onto 2 mm thick SDS-PAGE gels poured without a comb. After electrophoresis the gel was surface stained in aqueous COOMASSIE™ Blue, and the 47 kD and 31 kD bands cut out with a scalpel. (The word COOMASSIE is a trade mark.) Gel slices were electroeluted into dialysis bags in electrophoresis running buffer at 15 V for 24 hours, and the dialysate dialysed further against 0.1% SDS. Samples could be concentrated by lyophilisation.

Example 2

Amino-acid Sequence Data from Proteins

Protein samples (about 10 µg) were subjected to conventional N-terminal amino-acid sequencing. The 47 kD and 31 kD polypeptides were N-terminally blocked, so cyanogen bromide peptides of the 47 kD and 31 kD peptides were

9

1 prepared, and some amino-acid sequence was derived from these. Cyanogen
2 bromide cleaves polypeptide chains at methionine residues, and thus cleaved the
3 47 kD and 31 kD polypeptides gave rise to 24 kD and 17 kD peptides. In
4 addition the 47 kD polypeptide gave a 20 kD peptide. The 24 kD and 17 kD
5 peptides had the same 9 N-terminal amino-acid residues. This fact, combined
6 with the obvious one that the 31 kD could not contain both peptides
7 consecutively, suggested that the 24 kD peptide arose for a partial digest, where
8 full digestion would yield the 17 kD peptide. The other striking conclusion is
9 that the 47 kD and 31 kD proteins are related, and the 31 kD could be a further
10 processed form of the 47 kD. The 9 amino-acid sequence was used to construct
11 an oligonucleotide probe for the 47 kD/31 kD gene(s).

12

13 Example 3

14

15 *Raising Antibodies to the 47 kD and 31 kD Polypeptides*

16

17 Polyclonal antibodies were prepared using the methodology of Catty and
18 Raykundalia (1988). The serum was aliquoted into 1 ml fractions and stored at
19 -20°C.

20

21 *Characterising Antibodies to the 47 kD and 31 kD Polypeptides*

22

23 Serum was immediately characterised using the Ochterloney double-diffusion
24 technique, whereby antigen and antibody are allowed to diffuse towards one
25 another from wells cut in agarose in borate-saline buffer. Precipitin lines are
26 formed where the two interact if the antibody 'recognises' the antigen. This test
27 showed that antibodies to both antigens had been formed, and furthermore that
28 extensive cross-reaction took place between the 47 kD and 31 kD polypeptides
29 and their respective antibodies. This is further indication that the 47 kD and 31
30 kD polypeptides are closely related, as suggested by their cyanogen bromide
31 cleavage patterns.

32

33

The gamma-globulin fraction of the serum was partially purified by precipitation with 50% ammonium sulphate, solubilisation in phosphate-buffered saline (PBS) and chromatography on a DE 52 cellulose ion-exchange column as described by Hill, 1984. Fractions containing gamma-globulin were monitored at 280 nm (OD₂₈₀ of 1.4 is equivalent to 1 mg/ml gamma-globulin) and stored at -20°C. The effective titre of the antibodies was measured using an enzyme-linked immunosorbant assay (ELISA). The wells of a polystyrene microtitre plate were coated with antigen (10-1000 ng) overnight at 4°C in carbonate coating buffer. Wells were washed in PBS-TWEEN™ and the test gamma globulin added at concentrations of 10, 1 and 0.1 µg/ml (approximately 1:100, 1:1000 and 1:10,000 dilutions). (The word TWEEN is a trade mark.) The diluent was PBS-Tween containing 2% polyvinyl pyrrolidone (PVP) and 0.2% BSA. Controls were preimmune serum from the same animal. Binding took place at 37°C for 3-4 hours. The wells were washed as above and secondary antibody (goat anti-rabbit IgG conjugated to alkaline phosphatase) added at a concentration of 1 µg/ml, using the same conditions as the primary antibody. The wells are again washed, and alkaline phosphatase substrate (p-nitrophenyl phosphate; 0.6 mg/ml in diethanol-amine buffer pH 9.8) added. The yellow colour, indicating a positive reaction, was allowed to develop for 30 minutes and the reaction stopped with 3M NaOH. The colour is quantified at 405 nm. More detail of this method is given in Hill, 1984. The method confirmed that the antibodies all had a high titre and could be used at 1 µg/ml concentration.

Example 4

Isolation of Total RNA from Immature Cocoa Beans

The starting material for RNA which should contain a high proportion of mRNA specific for the storage proteins was immature cocoa beans, at about 130 days after pollination. Previous work had suggested that synthesis of storage proteins was approaching its height by this date (Biehl *et al*, 1982). The beans are roughly corrugated and pale pinkish-purple at this age.

1
2 The initial requirement of the total RNA preparation from cocoa beans was that
3 it should be free from contaminants, as judged by the UV spectrum, particularly
4 in the far UV, where a deep trough at 230 nm (260 nm : 230 nm ratio is
5 approximately 2.0) is highly diagnostic of clean RNA, and is intact, as judged
6 by agarose gel electrophoresis of heat-denatured samples, which should show
7 clear rRNA bands. A prerequisite for obtaining intact RNA is scrupulous
8 cleanliness and rigorous precautions against RNases, which are ubiquitous and
9 extremely stable enzymes. Glassware is customarily baked at high
10 temperatures, and solutions and apparatus treated with the RNase inhibitor
11 diethyl pyrocarbonate (DEPC, 0.1%) before autoclaving.
12

13 The most routine method for extraction of plant (and animal) RNA is extraction
14 of the proteins with phenol/chloroform in the presence of SDS to disrupt
15 protein-nucleic acid complexes, and inhibit the RNases which are abundant in
16 plant material. Following phenol extraction the RNA is pelleted on a caesium
17 chloride gradient before or after ethanol precipitation. This method produced
18 more or less intact RNA, but it was heavily contaminated with dark brown
19 pigment, probably oxidised polyphenols and tannins, which always co-purified
20 with the RNA. High levels of polyphenols are a major problem in *Theobroma*
21 tissues.
22

23 A method was therefore adopted which avoided the use of phenol, and instead
24 used the method of Hall *et al.* (1978) which involves breaking the tissue in hot
25 SDS-borate buffer, digesting the proteins with proteinase K, and specifically
26 precipitating the RNA with LiCl. This method gave high yields of reasonably
27 clean, intact RNA. Contaminants continued to be a problem and the method
28 was modified by introducing repeated LiCl precipitation steps, the precipitate
29 being dissolved in water and clarified by microcentrifugation after each step.
30 This resulted in RNA preparations with ideal spectra, which performed well in
31 subsequent functional tests such as *in vitro* translation.
32
33

Preparation of mRNA From Total RNA

The mRNA fraction was separated from total RNA by affinity chromatography on a small (1 ml) oligo-dT column, the mRNA binding to the column by its poly A tail. The RNA (1-2 mg) was denatured by heating at 65°C and applied to the column in a high salt buffer. Poly A+ was eluted with low salt buffer, and collected by ethanol precipitation. The method is essentially that of Aviv and Leder (1972), modified by Maniatis *et al* (1982). From 1 mg of total RNA, approximately 10-20 µg polyA+ RNA was obtained (1-2%).

In vitro Translation of mRNA

The ability of mRNA to support *in vitro* translation is a good indication of its cleanliness and intactness. Only mRNAs with an intact polyA tail (3' end) will be selected by the oligo-dT column, and only mRNAs which also have an intact 5' end (translational start) will translate efficiently. *In vitro* translation was carried out using RNA-depleted wheat-germ lysate (Amersham International), the *de novo* protein synthesis being monitored by the incorporation of [³⁵S]-methionine (Roberts and Paterson, 1973). Initially the rate of *de novo* synthesis was measured by the incorporation of [³⁵S]-methionine into TCA-precipitable material trapped on glass fibre filters (GFC, WHATMAN[™]). (The word WHATMAN is a trade mark.) The actual products of translation were investigated by running on SDS-PAGE, soaking the gel in fluor, drying the gel and autoradiography. The mRNA preparations translated efficiently and the products covered a wide range of molecular weights, showing that intact mRNAs for even the largest proteins had been obtained. None of the major translation products corresponded in size to the 47kD or 31kD storage polypeptides identified in mature beans, and it was apparent that considerable processing of the nascent polypeptides must occur to give the mature forms.

Example 5

Identification of Precursor to the 47 kD and 31 kD Polypeptides by Immunoprecipitation

5

10

Because the 47 kD and 31 kD storage polypeptides were not apparent amongst the translation products of mRNA from developing cocoa beans, the technique of immunoprecipitation, with specific antibodies raised to the storage polypeptides, was used to identify the precursors from the translation mixture. This was done for two reasons: first to confirm that the appropriate mRNA was present before cloning, and second to gain information on the expected size of the encoding genes.

15

20

Immunoprecipitation was by the method of Cuming *et al*, 1986. [³⁵S]-labelled *in vitro* translation products were dissociated in SDS, and allowed to bind with specific antibody in PBS plus 1% BSA. The antibody-antigen mixture was then mixed with protein A-SEPHAROSE™ and incubated on ice to allow the IgG to bind to protein A. (The word SEPHAROSE is a trade mark.) The slurry was poured into a disposable 1 ml syringe, and unbound proteins removed by washing with PBS +1% NONIDET™ P-40. (The word NONIDET is a trade mark.) The bound antibody was eluted with 1M acetic acid and the proteins precipitated with TCA. The antibody-antigen complex was dissociated in SDS, and subject to SDS-PAGE and fluorography, which reveals which labelled antigens have bound to the specific antibodies.

25

30

The results showed that the anti-47 kD and anti-31 kD antibodies both precipitated a 67 kD precursor. The precursor size corresponded to a major band on the *in vitro* translation products. The results with the 47 kD and 31 kD antibodies confirmed that the polypeptides are derived from a single precursor, or at least precursors of the same size. The large size of the precursor suggested that size-selection at mRNA or cDNA level may be necessary to obtain clones.

Example 6

cDNA Synthesis From the mRNA Preparations

cDNA synthesis was carried out using a kit from Amersham International. The first strand of the cDNA is synthesised by the enzyme reverse transcriptase, using the four nucleotide bases found in DNA (dATP, dTTP, dGTP, dCTP) and an oligo-dT primer. The second strand synthesis was by the method of Gubler and Hoffman (1983), whereby the RNA strand is nicked in many positions by RNase H, and the remaining fragments used to prime the replacement synthesis of a new DNA strand directed by the enzyme *E. coli* DNA polymerase I. Any 3' overhanging ends of DNA are filled in using the enzyme T4 polymerase. The whole process was monitored by adding a small proportion of [³²P]-dCTP into the initial nucleotide mixture, and measuring the percentage incorporation of label into DNA. Assuming that cold nucleotides are incorporated at the same rate, and that the four bases are incorporated equally, an estimate of the synthesis of cDNA can be obtained. From 1 µg of mRNA approximately 140 ng of cDNA was synthesised. The products were analysed on an alkaline 1.4% agarose gel as described in the Amersham methods. Globin cDNA, synthesised as a control with the kit, was run on the same gel, which was dried down and autoradiographed. The cocoa cDNA had a range of molecular weights, with a substantial amount larger than the 600 bp of the globin cDNA.

Example 7

Cloning of cDNA into a Plasmid Vector by Homopolymer Tailing

The method of cloning cDNA into a plasmid vector was to 3' tail the cDNA with dC residues using the enzyme terminal transferase (Boehringer Corporation Ltd), and anneal into a *Pst*I-cut and 5' tailed plasmid (Maniatis *et al.*, 1982; Eschenfeldt *et al.*, 1987). The optimum length for the dC tail is 12-20 residues. The tailing reaction (conditions as described by the manufacturers) was tested

1 with a 1.5 kb blunt-ended restriction fragment, taking samples at intervals, and
2 monitoring the incorporation of a small amount of [³²P]-dCTP. A sample of
3 cDNA (70 ng) was then tailed using the predetermined conditions.

4
5 A dG-tailed plasmid vector (3'-oligo(dG)-tailed pUC9) was purchased from
6 Pharmacia. 15 ng vector was annealed with 0.5 - 5 ng of cDNA at 58°C for 2
7 hours in annealing buffer: 5mM Tris-HCl pH 7.6; 1mM EDTA, 75 mM NaCl
8 in a total volume of 50 µl. The annealed mixture was transformed into *E. coli*
9 RRI (Bethesda Research Laboratories), transformants being selected on L-agar
10 + 100 µg/ml ampicillin. Approximately 200 transformants per ng of cDNA
11 were obtained. Transformants were stored by growing in 100 µl L-broth in the
12 wells of microtitre plates. adding 100 µl 80% glycerol, and storing at -20°C.

13
14 Some of the dC tailed cDNA was size selected by electrophoresing on a 0.8%
15 agarose gel, cutting slits in the gel at positions corresponding to 0.5, 1.0 and
16 1.5 kb, inserting DE81 paper and continuing electrophoresis until the cDNA
17 had run onto the DE81 paper. The DNA was then eluted from the paper with
18 high salt buffer, according to the method of Dretzen *et al* (1981).

19 20 21 Example 8

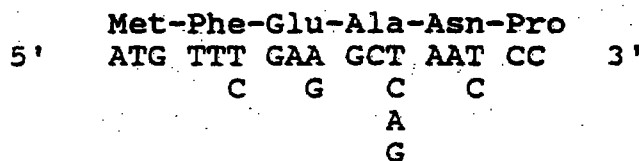
22 23 *Construction of Oligonucleotide Probes for the 47/31 kD Gene*

24
25 The amino-acid sequence obtained from a cyanogen bromide peptide common to
26 the 47 kD and 31 kD polypeptides is as follows:

27
28 Met-Phe-Glu-Ala-Asn-Pro-Asn-Thr-Phe

29
30 and the least redundant probe of 17 residues (a mixture of 32) is shown below:

31
32
33



The actual probe was made anti-sense so that it could also be used to probe mRNA. Probe synthesis was carried out using an Applied Biosystems apparatus.

Example 9

Use of Oligonucleotides to Probe cDNA Library

The oligonucleotide probes were 5' end-labelled with gamma-[³²P] dATP and the enzyme polynucleotide kinase (Amersham International). The method was essentially that of Woods (1982, 1984), except that a smaller amount of isotope (15 µCi) was used to label about 40 ng probe, in 10 mM MgCl₂, 100 mM Tris-HCl, pH 7.6; 20 mM 2-mercaptoethanol.

The cDNA library was grown on GeneScreen (New England Nuclear) nylon membranes placed on the surface of L-agar + 100 µg/ml ampicillin plates. (The word GeneScreen is a trade mark.) Colonies were transferred from microtitre plates to the membranes using a 6 x 8 multi-pronged device, designed to fit into the wells of half the microtitre plate. Colonies were grown overnight at 37°C, lysed in sodium hydroxide and bound to membranes as described by Woods (1982, 1984). After drying the membranes were washed extensively in 3 x SSC/0.1% SDS at 65°C, and hybridised to the labelled probe, using a HYBAID apparatus from Hybaid Ltd, PO Box 82, Twickenham, Middlesex. (The word HYBAID is a trade mark.) Conditions for hybridisation were as described by Mason & Williams (1985), a T_d being calculated for each oligonucleotide according to the formula:

$T_d = 4^\circ\text{C per GC base pair} + 2^\circ\text{C per AT base pair.}$

At mixed positions the lowest value is taken.

Hybridisation was carried out at $T_d - 5^\circ\text{C}$. Washing was in 6 x SSC, 0.1% SDS initially at room temperature in the HYBAID apparatus, then at the hybridisation temperature ($T_d - 5^\circ\text{C}$) for some hours, and finally at T_d for exactly 2 minutes. Membranes were autoradiographed onto FUJI X-ray film, with intensifying screens at -70°C . (The word FUJI is a trade mark.) After 24 - 48 hours positive colonies stood out as intense spots against a low background.

Example 10

Analysis of Positive Clones for the 47 kD/31 kD Polypeptide

Only one positive clone, pMS600, was obtained. This released two *Pst*I fragments on digestion, of total length 1.3 kb, insufficient to encode the 67 kD precursor. The total insert was removed from the vector on a *Hind*III-*Eco*RI fragment, nick-translated and used to probe the cDNA library, picking up a further two positive clones, pMS700 and pMS800. Restriction mapping of all three inserts suggested an overlapping map covering nearly 2.0 kb, sufficient to encode the 67 kD precursor (see Figure 1).

Example 11

Sequencing the Cloned Inserts

The sequencing strategy was to clone the inserts, and where appropriate subclones thereof, into the multiple cloning site of the plasmids pTZ18R/pTZ19R (Pharmacia). These plasmids are based on the better-known vectors pUC18/19 (Norrander *et al.* 1983), but contain a single-stranded origin of replication from the filamentous phage f1. When superinfected with phages in the same group, the plasmid is induced to undergo single-stranded

1 replication, and the single-strands are packaged as phages extruded into the
2 medium. DNA can be prepared from these 'phages' using established methods
3 for M13 phages (Miller, 1987), and used for sequencing by the method of
4 Sanger (1977) using the reverse sequencing primer. The superinfecting phage
5 used is a derivative of M13 termed M13K07, which replicates poorly and so
6 does not compete well with the plasmid, and contains a selectable
7 kanamycin-resistance marker. Detailed methods for preparing single-strands
8 from the pTZ plasmids and helper phages are supplied by Pharmacia. DNA
9 sequence was compiled and analysed using the Staden package of programs
10 (Staden, 1986), on a PRIME 9955 computer. (The word PRIME is a trade
11 mark.)

12

13 Example 12

14

15 *Features of the 47 kD/31 kD cDNA and Deduced Amino-acid Sequence of the 67* 16 *kD Precursor*

17

18 DNA sequencing of the three positive clones, pMS600, pMS700, pMS800,
19 confirmed the overlap presumed in Figure 1. No sequence differences were
20 found in the overlapping regions (about 300 bp altogether), suggesting that the
21 three cDNAs were derived from the same gene. The sequence of the combined
22 cDNAs comprising 1818 bases is shown in Figure 2. The first ATG codon is
23 found at position 14, and is followed by an open reading frame of 566 codons.
24 There is a 104-base 3' untranslated region containing a polyadenylation signal at
25 position 1764. The oligonucleotide probe sequence is found at position 569.

26

27 The open reading frame translates to give a polypeptide of 566 amino-acids
28 (Figure 2), and a molecular weight of 65612, which is reasonably close to the
29 67 kD measured on SDS-PAGE gels. The N-terminal residues are clearly
30 hydrophobic and look like a characteristic signal sequence. Applying the rules
31 of Von Heijne (1983), which predict cleavage sites for signal sequences, suggests
32 a cleavage point between amino-acids 20 and 21 (see Figure 3). The region
33 following this is highly hydrophobic and contains four Cys-X-X-X-Cys motifs.

19

1 The N-terminus of the mature protein has been roughly identified as the
2 glutamate (E) residue at 135 (Figure 3), on the basis of some tentative
3 N-terminal sequence (EEPGSQFANPAYHF). This N-terminus would give a
4 mature protein of 49068 kD, in rough agreement with that observed. There
5 appears to be no glycosylation sites (Asn-X-Ser/Thr) in the mature protein of
6 the sequence.

7 8 *Homologies Between the 67 kD Precursor and Other Known Proteins*

9
10 Searches through the PIR database, and through the literature, revealed close
11 homologies between the 67 kD polypeptide and a class of seed storage proteins
12 termed vicilins, one of two major classes of globulins found in seed (Borrito
13 and Dure, 1987). Alignments between the 67 kD polypeptide and vicilins from
14 cotton (*Gossypium hirsutum*, Ghi), soybean (*Glycine max*, Gma), pea (*Pisum*
15 *sativum*, Psa-c is convicilin, Psa-v is vicilin) and bean (*Phaseolus vulgaris*, Pvu)
16 are shown in Figure 4 (Bown *et al*, 1988; Chlan *et al*, 1986; Doyle *et al*, 1986;
17 Lycett *et al*, 1983). Identical residues are boxed.

18
19 All the vicilins have a mature molecular weight of around 47 kD, with the
20 exception of soybean conglycinin alpha and alpha1 subunits, which are 72 kD
21 and 76 kD respectively, and pea convicilin with a mature molecular weight of
22 64kD. The pea and bean subunits (2 sub classes each) are synthesised as small
23 precursors, around 50 kD. The most striking homology with the 67 kD is the
24 cotton vicilin (Chlan *et al*, 1986). Cotton is also the most closely related to
25 cocoa: both are members of the order Malvales. Interestingly cotton also has a
26 large precursor, of 69 kD, comprising a short signal sequence, a large
27 hydrophilic domain containing six Cys-X-X-X-Cys motifs, and a mature
28 domain. It may therefore be possible to synthesise the corresponding cotton
29 protein, by techniques analogous to those disclosed in this application and to use
30 the cotton protein, or fragments of it, in the preparation of flavour components
31 analogous to cocoa flavour components.

32

33

1 Example 13

3 *Expression of the 67 kD Polypeptide in E. coli*

5 Before the 67 kD coding region could be inserted into a expression vector the
6 overlapping fragments from the three separate positive clones had to be spliced
7 into a continuous DNA segment. The method of splicing is illustrated in Figure
8 6: a *HindIII*-*BglIII* fragment from pMS600, a *BglIII*-*EcoRI* fragment from
9 pMS700 and an *EcoRI*-*Sall* fragment from pMS800 were ligated into pTZ19R
10 cut with *HindIII* and *Sall*. The resulting plasmid, containing the entire 67 kD
11 cDNA, was termed pMS900.

13 An *NcoI* site was introduced at the ATG start codon, using the mutagenic
14 primer:

16 5' TAG CAA CCA TGG TGA TCA 3'.

18 *In vitro* mutagenesis was carried out using a kit marketed by Amersham
19 International, which used the method of Eckstein and co-workers (Taylor *et al*,
20 1985). After annealing the mutagenic primer to single-stranded DNA the
21 second strand synthesis incorporates alpha-thio-dCTP in place of dCTP. After
22 extension and ligation to form closed circles, the plasmid is digested with *NciI*,
23 an enzyme which cannot nick DNA containing thio-dC. Thus only the original
24 strand is nicked, and subsequently digested with exonuclease III. The original
25 strand is then resynthesised, primed by the remaining DNA fragments and
26 complementing the mutated position in the original strand. Plasmids are then
27 transformed into *E. coli* and checked by plasmid mini preparations.

29 The 67 kD cDNA was then cloned into the *E. coli* expression plasmid, pJLA502
30 (Figure 5), on an *NcoI* - *Sall* fragment (pMS902).

1 pJLA502 (Schauder *et al.* 1987) is marketed by Medac GmbH, Postfach
2 303629, D-7000, Hamburg 36 and contains the strong lambda promoters, P_L
3 and P_R , and the leader sequence and ribosome binding site of the very
4 efficiently translated *E. coli* gene, *atpE*. It also contains a temperature-sensitive
5 cI repressor, and so expression is repressed at 30°C and activated at 42°C. The
6 vector has an *NcoI* site (containing an ATG codon: CCATGG) correctly placed
7 with respect to the ribosome binding site, and foreign coding sequences must be
8 spliced in at this point.

9
10 The expression vector was transformed into *E. coli* UT580. The transformed
11 strain was grown in L-broth + ampicillin (100 µg/ml) at 30°C until log phase
12 ($OD_{610} = 0.5$) and the temperature was then shifted to 42°C and samples taken
13 at intervals. Samples were dissociated by boiling in SDS loading buffer, and
14 run on SDS-PAGE gels. The proteins were electroblotted onto nitrocellulose
15 membranes (Towbin *et al.*, 1979) and Western blotting carried out using the
16 anti-21 kD antibody prepared in Example 3 above (at 2 µg/ml) and as a
17 secondary antibody, goat anti-rabbit-IgG conjugated to alkaline phosphatase
18 (Scott *et al.*, 1988).

19
20 A specific band at 67 kD was produced by pMS902, showing that a functional
21 gene was present.

22 23 Example 14

24 25 *Expression of the 67 kD Polypeptide in Yeast*

26
27 Two yeast expression vectors were used, both based on a yeast-*E. coli* shuttle
28 vector containing yeast and *E. coli* origins of replication, and suitable selectable
29 markers (ampicillin-resistance for *E. coli* and leucine auxotrophy for yeast).
30 Both vectors contain the yeast pyruvate kinase (PK) promoter and leader
31 sequence and have a *HindIII* cloning site downstream of the promoter. One
32 vector, A (YVA), is designed for internal expression, and the other, B (YVB),
33 for secreted expression, having a portion of the signal sequence of the yeast

1 mating alpha-factor downstream of the promoter, with a *HindIII* site within it to
2 create fusion proteins with incoming coding sequences. The vectors are
3 illustrated in Figure 7.

4
5 To use the vectors effectively it is desirable to introduce the foreign coding
6 region such that for vector A, the region from the *HindIII* cloning site to the
7 ATG start is the same as the yeast PK gene, and for vector B, the remainder of
8 the alpha-factor signal, including the lysine at the cleavage point. In practice
9 this situation was achieved by synthesising two sets of *HindIII* - *NcoI* linkers to
10 breach the gap between the *HindIII* cloning site in the vector and the *NcoI* at the
11 ATG start of the coding sequence. This is illustrated in Figure 8.

12
13 In order to use the yeast vector B, the hydrophobic signal sequence must first be
14 deleted from the 67 kD cDNA. Although direct evidence of the location of the
15 natural cleavage site was lacking, the algorithm of Von Heijne predicts a site
16 between amino-acids 20 (alanine) and 21 (leucine). However it was decided to
17 remove amino-acids 2-19 by deletion, so that the useful *NcoI* site at the
18 translation start would be maintained.

19
20
21 For ease of construction of the yeast vectors, the strategy was to first clone the
22 *HindIII* - *NcoI* linkers into the appropriate pTZ plasmids, and then to clone the
23 linkers plus coding region into the yeast vectors on *HindIII* - *BamHI* fragments.
24 However the coding region contains an internal *BamHI* which must be removed
25 by in vitro mutagenesis, giving a new plasmid pMS903. The signal sequence
26 was deleted from pMS903 using the mutagenic primer

27
28 5' AGCATAGCAACCATGGTTGCTTTGTTCT 3'

29
30 to give pMS904. The appropriate *HindIII* - *NcoI* linkers were then cloned into
31 pMS903 and pMS904 to give pMS907 and pMS905 respectively, and the
32 *HindIII* - *BamHI* fragments (linkers + coding region) subcloned from these
33

1 intermediate plasmids into YVA and YVB respectively to give the yeast
2 expression plasmids pMS908 and pMS906. A diagrammatic scheme for these
3 and other constructs is given in Figure 9.

4
5 Because the mature cocoa protein appears to lack the N-terminal hydrophilic
6 domain, as described in Example 12, expression vectors have also been
7 designed to express the mature protein directly. Yeast is unlikely to have the
8 same processing enzymes as cocoa and optimum expression may be obtained for
9 a protein as close as possible to that found naturally in cocoa. Hence the DNA
10 encoding the hydrophilic domain (amino acids 20-134) was deleted from the
11 intermediate plasmids pMS907 and pMS905 to give plasmids pMS911 and
12 pMS909 respectively, and the *Hind*III - *Bam*HI fragments for these were cloned
13 into YVA and YVB to give the expression plasmids pMS912 and pMS910
14 (Figure 9).

15
16 A further modification was introduced by constructing expression in which the
17 plant terminator had been removed and replaced with the yeast ADH terminator
18 (present in YVA and YVB). The plant signal was removed by cutting the
19 intermediate plasmids pMS907 and pMS905 at the *Pvu*II site immediately
20 downstream of the coding region, at position 1716 in Figure 2. *Hind*III linkers
21 were added and the entire coding region cloned into the yeast expression vectors
22 on *Hind*III - *Hind*III fragments giving expression plasmids pMS914 (YVA) and
23 pMS916 (YVB) (Figure 9). A summary of the constructs made is given in
24 Figure 10.

25
26 The yeast expression plasmids were transferred into yeast spheroplasts using the
27 method of Johnston (1988). The transformation host was the LEU⁻ strain
28 AH22, and transformants were selected on leucine-minus minimal medium.
29 LEU⁺ transformants were streaked to single colonies, which were grown in 50
30 ml YEPD medium (Johnston, 1988) at 28 °C for testing the extent and
31 distribution of foreign protein. Cells were harvested from cultures in
32 preweighed tubes in a bench-top centrifuge, and washed in 10 ml lysis buffer
33 (200 mM Tris, pH 8.1; 10% glycerol). The cell medium was reserved and

1 concentrated 10-25 x in an AMICON mini concentrator. (The word AMICON
2 is a trade mark.) The washed cells were weighed and resuspended in lysis
3 buffer plus protease inhibitors (1 mM phenyl methyl sulphonyl fluoride
4 (PMSF); 1 μ g/ml aprotinin; 0.5 μ g/ml leupeptin) at a concentration of 1 g/ml.
5 1 volume acid-washed glass-beads was added and the cells broken by vortexing
6 for 8 minutes in total, in 1 minute bursts, with 1 minute intervals on ice. After
7 checking under the microscope for cell breakage, the mixture was centrifuged at
8 7000 rpm for 3 minutes to pellet the glass beads. The supernatant was removed
9 to a pre-chilled centrifuge tube, and centrifuged for 1 hour at 20,000 rpm.
10 (Small samples can be centrifuged in a microcentrifuge in the cold.) The
11 supernatant constitutes the soluble fraction. The pellet was resuspended in 1 ml
12 lysis buffer plus 10% SDS and 1% mercaptoethanol and heated at 90°C for 10
13 minutes. After centrifuging for 15 minutes in a microcentrifuge the supernatant
14 constitutes the particulate fraction.

15
16 Samples of each fraction and the concentrated medium were examined by
17 Western blotting. Considering first the plasmids designed for internal
18 expression in YVA, pMS908 produced immunoreactive proteins at 67 kD and
19 16 kD within the cells only. There was no evidence of the 67 kD protein being
20 secreted under the influence of its own signal sequence. The smaller protein is
21 presumed to be a degradation product. A similar result, but with improved
22 expression, was obtained with pMS914, in which the plant terminator is
23 replaced by a yeast terminator. However in pMS912, in which the coding
24 region for the hydrophilic domain has been deleted, no synthesis of
25 immunoreactive protein occurred.

26
27 For industrial production of heterologous proteins in yeast a secreted mode is
28 preferable because yeast cells are very difficult to disrupt, and downstream
29 processing from total cell protein is not easy. The results from the vectors
30 constructed for secreted expressed were rather complicated. From the simplest
31 construct, pMS906, in which the yeast α -factor signal sequence replaces the
32 plant protein's own-signal, immunoreactive proteins of approximately 47 kD, 28
33 kD and 18-20 kD were obtained and secreted into the medium. At first sight

25

1 this is surprising because the coding region introduced should synthesise a 67
2 kD protein. However the most likely explanation is that the yeast's KEX2
3 protease, that recognises and cleaves the α -factor signal at a Lys-Arg site is also
4 cleaving the 67 KD protein at Lys-Arg dipeptides at positions 148 and 313 in
5 the amino-acid sequence. The calculated protein fragment sizes resulting from
6 cleavage at these positions are 47179 Daltons, 28344 Daltons and 18835
7 Daltons, very close to the observed sizes.

8
9 When the plant terminator is replaced with a yeast terminator in pMS916 no
10 expression is obtained in either cells or medium; it is possible that a mutation
11 has been inadvertently introduced. From the construct pMS910, in which the
12 hydrophilic domain has been deleted the main antigenic products were 28 KD
13 and 18-20 kD, again secreted into the medium. It is presumed that the *de novo*
14 47 kD product is immediately cleaved at the KEX2 site at position 313.

15
16 In summary, four of the six expression vectors constructed direct the synthesis
17 of proteins cross-reacting with anti-47 kD antibodies. Two of the constructs
18 secrete proteins into the medium.

19 20 Example 15

21 22 *Construction of Vectors Designed to Express the 67 kD Protein in Hansenula* 23 *polymorpha*

24
25 The methylotropic yeast *Hansenula polymorpha* offers a number of advantages
26 over *Saccharomyces cerevisiae* as a host for the expression of heterologous
27 proteins (EP-A-0173378 and Sudbery *et al.*, 1988). The yeast will grow on
28 methanol as sole carbon source, and under these conditions the enzyme
29 methanol oxidase (MOX) can represent up to 40% of the total cell protein.
30 Thus the MOX promoter is a very powerful one that can be used in a vector to
31 drive the synthesis of heterologous proteins, and it is effective even as a single
32 copy. This gives the potential to use stable integrated vectors. *Hansenula* can
33 also grow on rich carbon sources such as glucose, in which case the MOX

promoter is completely repressed. This means that cells containing the heterologous gene can be grown to a high density on glucose, and induced to produce the foreign protein by allowing the glucose to run out and adding methanol.

A plasmid, pHGL1, containing the MOX promoter and terminator, and a cassette containing the yeast α -factor secretory signal sequence, were prepared. The 67 kD coding region was cloned into pHGL1 on a *Bam*HI - *Bam*HI fragment, replacing the *Bgl*III fragment which contains the 3' end of the MOX coding region. The whole promoter - gene - terminator region can then be transferred to YEp13 on a *Bam*HI - *Bam*HI fragment to give the expression plasmid pMS922. The details of the construction are illustrated in Figure 11. An analogous expression plasmid, pMS925, has been constructed with the yeast α -factor spliced onto the 67 kD coding region, replacing the natural plant signal. The *Bam*HI - *Hind*III cassette containing the α -factor was ligated to the *Hind*III - *Bam*HI fragment used to introduce the 67 kD coding region into YVB. The α -factor plus coding region was then cloned with pHGL1 on a *Bam*HI - *Bam*HI fragment, and transferred into YEP13 as before. Details are shown in Figure 12.

Both constructs have been transformed into *Hansenula* and grown under inducing conditions with 0.5% or 1% methanol. Both constructs directed the production of immunoreactive protein within the cells, and pMS925 secreted the protein into the medium under the influence of the α -factor signal sequence.

E. coli Strains

RR1	$F^{v_B} M_B$ <i>ara-14 proA2 leuB6 lacY1 galK2 vpsL20 (str^r)</i>
	<i>xyl-5 ml-1 supE44</i>
CAG629	<i>lac_{am} rvp_{am} pho_{am} htpR_{am} mal rpsL lon supC_{ts}</i>

1 UT580 (lac-pro) supE thi hsdD5 / F' tra D36 proA⁺B⁺ lacI^q lacZ
2 M15
3
4
5

6 References
7

8 Aviv, H., and Leder, P. *Proc. Natl. Acad. Sci. USA* 69, 1408-1412 (1972).
9 Purification of biologically active globin mRNA by chromatography on oligo
10 dT cellulose
11

12 Biehl, B., Wewetzer, C., and Passern, D. *J. Sci. Food Agric.* 33, 1291-1304
13 (1982). Vacuolar (Storage) Proteins of Cocoa Seeds and their Degradation
14 during Germination and Fermentation.
15

16 Borroto, K., and Dure, L. *Plant Molecular Biology* 8, 113-131 (1987). The
17 globulin seed storage proteins of flowering plants are derived from two ancestral
18 genes.
19

20 Bown, D., Ellis, T.H.N., and Gatehouse, J.A. *Biochem. J.* 251, 717-726
21 (1988). The sequence of a gene encoding convicilin from pea (*Pisum sativum*)
22 shows that convicilin differs from vicilin by an insertion near the N-terminus.
23

24 Catty, D. and Raykundalia, C. Production and Quality control of Polyclonal
25 Antibodies in: "Antibodies: A Practical Approach" Vol I, IRL Press (1988)
26

27 Chlan, C.A., Pyle, J.B., Legocki, A.B., and Dure, L. *Plant Molecular*
28 *Biology* 7, 475-489 (1986). Developmental biochemistry of cotton seed
29 embryogenesis and germination XVIII. cDNA and amino acid sequences of
30 members of the storage protein families.
31
32
33

- 1 Cuming, A.C., Williams, R.S., and Cullimore, J.V. in "Immunology in Plant
2 Science", Ed. Wang, T.L., Cambridge University Press, 1986. The use of
3 Antibodies in Molecular Biology.
4
- 5 Doyle, J.J., Schuler, M.A., Godette, W.D., Zencer, V., Beachy, R.N., and
6 Slightom, J.L. *J. Biol. Chem.* 261, 9228-9238 (1986). The glycosylated seed
7 storage proteins of Glycine max and Phaseolus vulgaris : Structural homologies
8 of genes and proteins.
9
- 10 Dretzen, G., Bellard, M., Sassone-Corsi, P., and Chambon, P. *Analytical*
11 *Biochemistry* 112, 295-298 (1981). A reliable method for the recovery of DNA
12 fragments from agarose and acrylamide gels.
13
- 14 Eschenfeldt, W.H., Puskas, R.S., and Berger, S.L. *Methods in Enzymology*
15 152, 337-342 (1987). Homopolymeric Tailing.
16
- 17 Fritz *et al* (*J. Food Sci.* 50 946-950 (1985))
18
- 19 Gubler, U., and Hoffman, B.J. *Gene* 25, 263 (1983). A simple and very
20 efficient method for generating cDNA libraries.
21
- 22 Hall, T.C., Ma, Y., Buchbinder, B.U., Pyne J.W., Sun, S.M., and Bliss,
23 F.A. *Proc. Natl. Acad. Sci. USA* 75, 3196-3200 (1978). Messenger RNA for
24 G1 protein of French bean seeds : cell-free translation and product
25 characterisation.
26
- 27 Hill, S.A. "Methods in Plant Virology", Blackwell 1984.
28
- 29 Johnston, J.R. in "Yeast : A practical approach". Eds Campbell, I., and
30 Duffus, J.H. IRL Press, 1988. Yeast Genetics. Molecular Aspects.
31
- 32 Kreil, G. *Annual Rev. Biochem.* 50, 317-348 (1981). Transfer of proteins
33 across membranes.

29

- 1
- 2 Laemmli, U.K. *Nature* 227, 680 (1970). Cleavage of structural proteins
- 3 during the assembly of the head of bacteriophage T4.
- 4
- 5 Lipman, D.J., and Pearson, W.R. *Science* 227, 1435-1441 (1985). Rapid
- 6 protein sequence similarity searches.
- 7
- 8 Lycett, G.W., Delauney, A.J., Gatehouse, J.A., Gilroy, J., Croy, R.R.D., and
- 9 Boulter, D. *Nucleic Acids Res.* 11, 2367-2380 (1983).
- 10
- 11 Maniatis, T., Fritsch, E.F., and Sambrook, J. "Molecular Cloning: A
- 12 Laboratory Manual", Cold Spring Harbour Laboratory, 1982.
- 13
- 14 Mason, P.J., and Williams, J.G. in "Nucleic Acid Hybridisation : A Practical
- 15 Approach". Ed. Hames, B.D., and Higgins, S.J. IRL Press 1985.
- 16 Hybridisation in the Analysis of Recombinant DNA.
- 17
- 18 Meinkoth, J., and Wahl, G.M. *Analytical Biochemistry* 138, 267 (1984).
- 19 Methods of Southern blotting and DNA probing.
- 20
- 21 Miller, H. *Methods in Enzymology* 152, 145-170 (1987). Practical Aspects of
- 22 Preparing Phage and Plasmid DNA : Growth, Maintenance and Storage of
- 23 Bacteria and Bacteriophage.
- 24
- 25 Norrander, J., Kempe, T., and Messing, J. *Gene* 26, 101 (1983).
- 26 Construction of improved M13 vectors using oligodeoxynucleotide-directed
- 27 mutagenesis.
- 28
- 29 Proudfoot, N.J., and Brownlee, G.G. *Nature* 263, 211-214 (1976). 3'
- 30 Non-coding region sequences in eukaryotic messenger RNA.
- 31
- 32
- 33

- 1 Roberts, B.E., and Paterson, B.M. *Proc. Natl. Acad. Sci. USA* 70, 2330
2 (1973). Efficient translation of tobacco mosaic virus RNA and rabbit globin 9S
3 RNA in a cell-free system from commercial wheat germ.
4
- 5 Sanger, F., Nicklen, S., and Coulson, A.R. *Proc. Natl. Acad. Sci. USA* 74,
6 5463-5467 (1977). DNA sequencing with chain-terminating inhibitors.
7
- 8 Schauder, B., Blocker, H., Frank, R., and McCarthy, J.E.G. *Gene* 52,
9 279-283 (1987). Inducible expression vectors incorporating the *E. coli* *aptE*
10 translation initiation region.
11
- 12 Scott, R., Draper, J., Jefferson, R., Dury, G., and Jacob, L. in "Plant Genetic
13 Transformation and Gene Expression: A Laboratory Manual". Eds. Draper,
14 J., Scott, R., Armitage, P., Walden, R. Blackwell 1988. Analysis of gene
15 organisation and expression in plants.
16
- 17 Staden, R. *Nucleic Acids Res.* 14, 217-231 (1986). The current status and
18 portability of our sequence handling software.
19
- 20 Sudbery, P.E., Gleeson, M.A., Veale, R.A., Lederboer, A.M., and
21 Zoetmulder, M.C.M. *Biocem. Soc. Trans* 16 1081-103 (1988). Hansenula
22 polymorpha as a novel yeast system for the expression of heterologous genes.
23
- 24 Taylor, J.W., Ott, J., and Eckstein, F. *Nucleic Acids Res.* 13, 8765-8785
25 (1985). The rapid generation of oligonucleotide-directed mutations at high
26 frequency using phosphorothioate-modified DNA.
27
- 28 Towbin, H., Staehelin, T., and Gordon, J. *Proc. Natl. Acad. Sci. USA* 76,
29 4350-4534 (1979). Electrophoretic transfer of proteins from polyacrylamide
30 gels to nitrocellulose sheets : procedure and some applications.
31
- 32 Von Heije, G. *Eur. J. Biochem* 133, 17-21 (1983). Patterns of Amino-acids
33 near Signal-Sequence Cleavage Sites.

- 1
- 2 Wahl, G.M., and Berger, S.L. *Methods in Enzymology* 152, 415-423 (1987).
- 3 Screening Colonies or Plaques with Radioactive Nucleic Acid Probes.
- 4 Woods, D.E. *Focus (Bethesda Research Labs)* 6, 3 (1984). Oligonucleotide
- 5 Screening of cDNA Libraries.
- 6
- 7 Woods, D.E., Markham, A.F., Ricker, A.T., Goldberger, G., and Colten,
- 8 H.R. *Proc. Natl. Acad. Sci. USA* 79, 5561 (1982).
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33



CLAIMS

1. A protein encoded by mRNA extractable from beans of cocoa (*Theobroma cacao*), the protein having a molecular weight of 67 kD, as assessed by SDS-PAGE, and at least part of the amino acid sequence shown in Figure 3 (including at least some of the unboxed residues in the sequence marked *Tca* in Figure 4), or a fragment of such a protein having at least 20 amino acids.
2. A protein extractable from beans of cocoa (*Theobroma cacao*), the protein having a molecular weight of 47 kD, as assessed by SDS-PAGE, and part of the amino acid sequence shown in Figure 3 (including at least some of the unboxed residues in the sequence marked *Tca* in Figure 4), or a fragment of such a protein having at least 20 amino acids.
3. A protein extractable from beans of cocoa (*Theobroma cacao*), the protein having a molecular weight of 31 kD, as assessed by SDS-PAGE, and part of the amino acid sequence shown in Figure 3 (including at least some of the unboxed residues in the sequence marked *Tca* in Figure 4), or a fragment of such a protein having at least 20 amino acids.
4. A protein as claimed in claim 1, 2 or 3 which is the product of recombinant DNA technology.
5. Recombinant or isolated nucleic acid encoding a protein or fragment as defined in claim 1, 2 or 3.
6. Recombinant or isolated nucleic acid which comprises:
 - (a) a nucleotide sequence as shown in Figure 2, encoding a 67, 47 or 31 kD protein (as measured by SDS-PAGE); or



33

- (b) a nucleotide sequence which hybridises to (a); or
- (c) a fragment of (a) or (b) of at least 60 nucleotides.

7. Nucleic acid as claimed in claim 6, which is DNA.

5

8. Nucleic acid as claimed in claim 7, which is a recombinant vector.

9. Nucleic acid as claimed in claim 8, wherein the vector is an expression vector and the protein- or fragment-coding sequence is operably linked to a promoter.

10

10. Nucleic acid as claimed in claim 9, wherein the expression vector is a yeast expression vector and the promoter is a yeast pyruvate kinase (PK) promoter.

15

11. Nucleic acid as claimed in claim 9, wherein the expression vector is a bacterial expression vector and the promoter is a strong λ promoter.

12. Nucleic acid as claimed in claim 9, 10 or 11, comprising a segment encoding a signal sequence.

20

13. A host cell comprising nucleic acid as claimed in claim 8.

14. A host cell comprising nucleic acid as claimed in any one of claims 9 to 12.

25

15. A host cell as claimed in claim 13 or 14, which is *Saccharomyces cerevisiae*.

16. A host cell as claimed in claim 13 or 14, which is *Hansenula polymorpha*.



34

17. A host cell as claimed in claim 13 or 14, which is *Escherichia coli*.

18. A process for the preparation of a protein or fragment as claimed in any one of claims 1 to 4, the process comprising coupling successive amino acids by peptide bond formation.

5

19. A process for the preparation of a protein or fragment as claimed in any one of claims 1 to 4, the process comprising culturing a host cell as claimed in any one of claims 14 to 17.

10

20. A process for the preparation of nucleic acid as claimed in any one of claims 5 to 12, the process comprising coupling together successive nucleotides and/or ligating oligo- or poly-nucleotides.
