Tems, Ursula (2009) ***Characterization of the carbonic anhydrase isozymes of Zea mays*****. PhD Thesis, James Cook University.**

Characterization of the Carbonic Anhydrase Isozymes of *Zea mays*

Thesis submitted by

Ursula TEMS B.Sc. Hons (JCU)

February 2009

for the degree of Doctor of Philosophy

in the School of Pharmacy and Molecular Sciences

James Cook University

Statement of Sources

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Signature Date

Statement of Access

I, the undersigned, author of this work, understand that James Cook University will make this thesis available for use within the University Library and, via the Australian Digital Theses network, for use elsewhere.

I understand that, as an unpublished work, a thesis has significant protection under the Copyright Act and;

I do not wish to place any further restriction on access to this work.

Signature Date

Declaration

I, the undersigned, the author of this work, declare that the electronic copy of this thesis provided to the James Cook University Library is an accurate copy of the print thesis submitted, within the limits of the technology available.

Signature Date

Acknowledgements

I would like to express my sincere gratitude to my supervisor Jim Burnell for his enthusiasm, encouragement and guidance.

I would like to acknowledge the work done by previous students from the Burnell Laboratory at James Cook University, which included the creation of several plasmid constructs that were used in this investigation. I am also grateful to Florigene Pty Ltd for the use of laboratory equipment and reagents.

Thank you to Florian, my family and my friends for their support, patience and understanding. I would like to also express thanks to my fellow students, particularly Griselda, who was the best lab mate imaginable.

Abstract

In maize, CA catalyzes the first reaction in the C_4 photosynthetic pathway, hydrating carbon dioxide that has diffused into the mesophyll cell cytoplasm to bicarbonate, providing an inorganic carbon source for the C_4 pathway. The beta-CA isozymes from maize, as well as other agronomically important C_4 crops such as sorghum and sugarcane, differ significantly from other reported forms of the enzyme and have remained relatively uncharacterized.

The mRNA transcripts encoding the CA isozymes contain repeating sequences of approximately 600 bp that encode multiple protein domains (Repeat A, Repeat B and Repeat C). In maize, three cDNA sequences had been determined and designated CA1, CA2 and CA3. There are at least three genes in the maize genome, and one of these encodes two identical protein domains, with distinct groups of exons corresponding to the repeating regions of the transcript. The first exon of the CA2 gene encodes a putative chloroplast transit peptide, indicating an additional non-photosynthetic role for CA in maize, such as in lipid biosynthesis pathways and/or replenishing the Krebs cycle intermediates together with PEP carboxylase. This is supported by the identification of CA transcripts in root tissue and analysis of the gene sequence, which identified promoter elements that direct constitutive expression.

The expression of a single repeat region of the transcript produced active enzyme, able to catalyze the reversible hydration of carbon dioxide to bicarbonate producing hydrogen ions. The carbon dioxide hydration activity of Repeat B was relatively high compared to the activity of either Repeat A or C. Repeat B was also found to be a dimer and is composed primarily of alpha-helices, in agreement with that observed for other plant CAs. The active site of the individual protein domains, Repeat A, Repeat B and Repeat C was identified and found to contain the conserved amino acids proposed to coordinate the catalytic zinc ion and act as a proton acceptor during regeneration of the active enzyme complex.

Table of Contents

Statement of Sources i
Statement of Access ii
Declaration........ ii
Acknowledgements iii
Abstract........ iv
Table of Contents v
List of Figures x
Abbreviations xix
Publications xxii

1. **Introduction** 1
1.1. CA gene families 2
1.1.1 Alpha CA gene family 2
1.1.2 Beta CA gene family 7
1.1.3 Gamma CA gene family 14
1.1.4 Delta CA gene family 14
1.1.5 Epsilon CA gene family 15
1.2. Photosynthesis 16
1.2.1 C_4 Photosynthesis 17
1.3. CA in C_3 plants 20
1.4. CA in C_4 plants 22
1.5. CA in NADP-ME type monocot C_4 plants 24
1.6. Summary 27

2. **Materials and Methods** 28
2.1. Molecular biology methods 28
2.1.1 Restriction endonuclease treatment 28
2.1.2 Agarose gel electrophoresis 28
2.1.3 Purification of DNA from agarose gels 28
2.1.4 Purification of total RNA from maize tissue 29
2.1.5 Determining the concentration of single and double stranded DNA, and RNA 29
2.1.6 Phenol/chloroform extraction and ethanol precipitation of DNA 29
2.1.7 Preparation of competent cells 30
2.1.8 Preparation of glycerol stocks 30
2.1.9 Transformation of competent cells and blue/white selection 30
2.1.10 Plasmid DNA preparation 31
2.1.11 Ligation reactions 32
2.1.12 Alkaline phosphatase treatment of DNA fragments 32
2.1.13 DNA Sequencing 32
2.1.14 Polymerase Chain Reaction (PCR) 33
2.2. Protein methods 34
2.2.1 Denaturing Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis (SDS-PAGE) 34
2.2.2 Western blotting 34

2.2.3 Ponceau staining 35
2.2.4 Immunoblotting 35
2.2.5 Bradford assay for protein concentration 35
2.3. Plant material 36

3. **Genomic Organization of the beta-CA Gene Family** 37
3.1. Introduction 37
3.1.1 Maize genome complexity 37
3.1.2 Sequencing of the maize genome 38
3.1.3 CA gene families 39
3.1.4 Rationale 41
3.2. Materials and methods 42
3.2.1 Purification of genomic DNA 42
3.2.2 Analysis of gene number by Southern blotting 43
3.2.5 Preparation of probes 44
3.2.6 Hybridization and detection 46
3.2.7 Calculation of the molecular weight of hybridizing fragments 47
3.2.8 Screening the maize genomic DNA library 47
3.2.9 Preparation of host cells 47
3.2.10 Genomic DNA library titration and plating phage 47
3.2.11 Genomic DNA library screening 48
3.2.12 Preparation of lambda DNA 49
3.2.13 Restriction endonuclease treatment of genomic DNA library clone 50
3.2.14 Southern hybridization of lambda DNA 50
3.2.15 Sequence analysis 51
3.2.16 Amplification of CA genomic DNA 51
3.3 Results 52
3.3.1 The maize CA2 gene 52
3.3.1.1 Sequence of the maize CA2 gene 52
3.3.1.2 Screening the maize genomic DNA library 53
3.3.1.3 Analysis of the genomic DNA library clone 53
3.3.2 Beta-CA multi-gene family 56
3.3.2.1 Determining gene number by Southern analysis 56
3.3.2.2 PCR analysis of maize genomic DNA 59
3.3.2.3 Interrogation of the Maize Genomic Databases 61
3.4 Discussion 68
3.4.1 The maize CA2 gene 68
3.4.1.1 Screening the maize genomic DNA library 68
3.4.1.2 Southern analysis of genomic DNA 69
3.4.2 Interrogation of the Maize Genomic Databases 72
3.4.3 The CA2 gene encodes two identical protein domains 73
3.5 Conclusion 75

4. **Analysis of the Maize CA Nucleotide Sequence** 77
4.1. Introduction 77
4.1.1 CA in C_4 monocot plant species 77
4.1.2 C_4 photosynthesis and NADP-ME catalyzed decarboxylation in bundle sheath cells…… 78
4.1.4 Rationale 82

4.2. Materials and methods 83
4.2.1 Generation of a phylogenetic tree 83
4.2.2 Analysis of the CA2 gene structure 83
4.2.2.1 Identification of intron/exon boundaries 83
4.2.2.2 Analysis of nucleotide composition of the CA gene in maize and rice 83
4.2.2.3 Identification of CpG islands 83
4.2.3 Transcription initiation site 84
4.2.3.1 Determining the transcription initiation site by 5′-rapid amplification of cDNA ends (5′-RACE) 84
4.2.3.2 PCR amplification of the 5′-end of the CA cDNA sequence 85
4.2.4 Analysis of transcription factor binding sites 86
4.2.5 Semi-quantitative reverse transcriptase PCR 86
4.3. Results 89
4.3.1 The CA2 gene of maize 89
4.3.1.1 Gene structure 89
4.3.1.2 Nucleotide composition 91
4.3.1.3 Identification of CpG islands 93
4.3.2 Transcription initiation 93
4.3.2.1 Identification of the consensus sequence 93
4.3.2.2 Transcription initiation site 94
4.3.3 Analysis of transcription regulation 96
4.3.4 Semi-quantitative RT-PCR 101
4.4. Discussion 105
4.4.1 Analysis of genomic sequence data 106
4.4.1.1 Intron/exon boundaries 106
4.4.1.2 Nucleotide composition 107
4.4.2 Transcription initiation site 108
4.4.3 Identification of putative regulatory elements 109
4.4.4 Analysis of transcript abundance in maize tissues 113
4.5. Conclusion 118

5. **Expression and Characterization of Maize CA** 120
5.1. Introduction 120
5.1.1 Molecular structure of maize CA 120
5.1.2 Enzymatic activity 122
5.1.3 Rational 123
5.2. Materials and methods 124
5.2.1 Source of CA plasmid constructs 124
5.2.2 Creation of CA expression constructs 124
5.2.2.1 RT-PCR primers and reactions 125
5.2.2.2 Restriction enzyme generation of constructs 126
5.2.2.3 Sequence analysis 127
5.2.3 Bacterial expression and purification 128
5.2.3.1 Inoculating cultures and inducing expression 128
5.2.3.2 Purification by affinity chromatography 128
5.2.4 Analysis of the expressed CA enzymes 129
5.2.4.1 SDS-PAGE and Western blots 129
5.2.4.2 Blot-affinity purification of antibodies 129
5.2.4.3 Antibody clearing 130

5.2.4.4 Secondary structure prediction based on the amino acid sequence 130
5.2.4.5 Measuring CA activity 130
5.2.4.6 CA inhibition analysis 131
5.3. Results 133
5.3.1 Expression of the CA enzymes 133
5.3.1.1 CA1 expression 134
5.3.1.2 Repeat A expression 134
5.3.1.3 Repeat B expression 135
5.3.1.4 Repeat C expression 136
5.3.1.5 CA4 expression 137
5.3.1.6 CA5 expression 137
5.3.1.7 CA6 expression 138
5.3.1.8 CA7 expression 139
5.3.1.9 CA8 expression 140
5.3.2 Analysis of the CA enzymes 141
5.3.2.1 Analysis of CA activity 141
5.3.2.2 Analysis of CA inhibition 143
5.4. Discussion 147
5.4.1 CA expression 147
5.4.2 Immuno-reactivity of the CA enzymes 149
5.4.3 Analysis of CA activity 150
5.4.4 Analysis of CA inhibition 154
5.5. Conclusion 157

6. **Analysis of the Maize CA Enzymes** 159
6.1. Introduction 159
6.1.1 CA in monocot and dicot plant species 159
6.1.2 Localization of CA in C_3 and C_4 plants 160
6.1.3 Structure of plant CA 162
6.1.4 Rationale 165
6.2. Materials and methods 166
6.2.1 Analysis of maize CA protein sequence 166
6.2.1.1 Identification of a putative chloroplast transit peptide in the CA2 gene 166
6.2.1.2 Analysis of CA hydrophobicity 166
6.2.2 Analysis of CA (Repeat B) secondary structure 166
6.2.2.1 Secondary structure prediction based on the amino acid sequence 166
6.2.2.2 Analysis of secondary structure by circular dichroism (CD) spectroscopy 167
6.2.2.3 Analysis of Repeat B by native PAGE 167
6.2.3 Analysis of the CA2 unique 276 bp insert sequence 168
6.2.3.1 Purification of maize leaf mRNA 168
6.2.3.2 Generation of cDNA 168
6.2.3.3 PCR 168
6.2.3.4 Sequence analysis 169
6.2.4 Analysis of CA from leaf tissue 169
6.2.5 Analysis of leaf proteins by Western blot 169
6.2.6 Gel filtration 170
6.2.6.1 Column calibration 170
6.2.6.2 Identification of fractions containing CA 171
6.2.6.3 ELISA 171

6.2.6.4 Trichloroacetic acid precipitation 172
6.2.7 Ammonium sulphate precipitation 172
6.2.8 Ion exchange chromatography 173
6.2.9 Immuno-precipitation 173
6.3. Results 174
6.3.1 Analysis of the CA2 gene sequence 174
6.3.1.1 Identification of a putative chloroplast transit peptide 174
6.3.1.2 Analysis of Hydrophobicity 176
6.3.2 Analysis of CA (Repeat B) secondary structure 177
6.3.2.1 Secondary structure prediction based on the amino acid sequence 177
6.3.2.2 Analysis of secondary structure by CD spectroscopy 178
6.3.2.3 Analysis of Repeat B by native PAGE 180
6.3.3 Analysis of the CA2 unique 276 bp insert sequence 180
6.3.4 Analysis of CA from leaf tissue 182
6.3.4.1 Analysis of maize leaf proteins by Western blot 183
6.3.4.2 Gel filtration 183
6.3.4.3 Ammonium sulphate precipitation and ion exchange chromatography 187
6.3.4.4 Immuno-precipitation 188
6.4. Discussion 189
6.4.1 Analysis of amino acid sequence data 189
6.4.1.1 Prediction of sub-cellular location 189
6.4.1.2 Hydrophobicity 190
6.4.1.3 Secondary structure predictions 191
6.4.2 Analysis of the unique 276 bp insert of CA2 192
6.4.3 Analysis of CA from maize leaf 193
6.5. Conclusion 195

7. **Conclusions** 197

References 200

Appendix - Chapter 3 217
Appendix - Chapter 4 247
Appendix - Chapter 5 254
Appendix - Chapter 6 259

List of Figures

Chapter 1

Fig 1.1. Reaction mechanism of CA showing (a) nucleophilic attack of the metal-activated hydroxide ion on carbon dioxide, (b) ligand exchange of the product bicarbonate for a water molecule, and (c) regeneration of the zinc hydroxide form of the enzyme (Cronk *et al*., 2001). 5

Fig. 1.2. Diagrammatic representation of the role of CA in the thylakoid lumen, supplying carbon dioxide to Rubisco as part of the carbon concentrating mechanism in *C. reinhardtii* (Park *et al*., 1999). 6

Fig. 1.3. Phylogenetic analysis of the beta-CA gene family (Hewett-Emmett and Tashian, 1996). 11

Fig 1.4. Amino acid alignment of beta-CA sequences (Roeske and Ogren, 1990; Majeau and Coleman, 1992; Suzuki and Burnell, 1995; Burnell *et al*., 1990a; Burnell and Ludwig, 1997). 13

Fig. 1.5. The overall reaction of photosynthesis, showing the conversion of carbon dioxide to carbohydrates, using water and light energy from the sun, while producing oxygen. 16

Fig. 1.6. Depiction of the C_4 photosynthetic pathway, showing anatomical compartmentalization of carbon assimilation and fixation that occurs in some C_4 plant species (Lea and Leegood, 1999). 17

Fig. 1.7. Schematic representation of the three beta-CA cDNAs from maize (Burnell and Ludwig, 1997; Burnell, unpublished). The repeat sequences are approximately 600 bp and are labelled A, B and C. The black shaded box represents the 276 bp insert unique to CA2. The dark grey shaded boxes represent the 5′-leader sequences and 3′-untranslated regions. 24

Chapter 3

Fig. 3.1. Schematic representation of the three beta-CA cDNAs from maize (Burnell and Ludwig, 1997; Burnell, unpublished). The repeat sequences are labelled A, B and C and are approximately 600 bp. The black shaded box represents the 276 bp insert unique to CA2. The dark grey shaded boxes represent the 5′-leader sequences and 3′-untranslated regions. 40

Fig. 3.2. (A) Schematic representation of the method for amplifying the probes used for hybridization with maize genomic DNA. (B) The primer combinations used to amplify the probes used. (C) The PCR products were visualized by agarose gel electrophoresis and ethidium bromide staining (see Appendix 3.1 for probe sequences). 45

Fig. 3.3. (A) Agarose gel electrophoresis of the *Eco*RI/*Xho*I treated CA2 plasmid construct. The 900 bp band was excised from the gel, purified and used to probe the maize genomic DNA library. (B) Schematic representation of the components of the CA2 probe sequence. The *Eco*RI recognition site was before the 5′-end of the cDNA sequence, in the multiple cloning site of the vector (see Appendix 3.2 for probe sequence). 49

Fig. 3.4. Schematic representation of the primer binding sites used to amplify CA genomic DNA and spanning both exons and introns. 51

Fig. 3.5. Schematic representation of the genomic 7.45 kb sequence that corresponded to the CA2 gene and contained the first nine exons of the CA2 cDNA sequence (GenBank accession U08401, version: gi:606810; Burnell and Ludwig, 1997). 52

Fig. 3.6. (A) Agarose gel electrophoresis of the genomic library clone after restriction endonuclease treatment with the restriction enzymes indicated. The molecular weight markers are in lanes 1 and 12. (B) Probing the restriction fragments with the 900 bp probe used to screen the genomic DNA library enabled identification of bands that contained sequence of interest, which could then be sub-cloned for further analysis. 54

Fig. 3.7. (A) Matrix alignment plot showing similarity between the CA2 gene sequence and the genomic library clone assembly. Similarity is based on a minimum alignment of 90% in a 30 bp window (hash value 6). (B) Schematic representation of the alignment of the assembled genomic library sequence, red, with the CA2 gene sequence. 55

Fig. 3.8. (A) Agaroge (0.8%) gel electrophoresis of 10 μg of maize genomic DNA; Southern blot analysis of maize genomic DNA (10 μg) (B) probed with the Repeat probe, and (C) probed with the Insert probe. (1) untreated, (2) treated with *Eco*RI and (3) treated with *Kpn*I. The molecular weight markers used are Hyperladder VI (Bioline) and λ*Hind*III (Promega). 57

Fig. 3.9. Standard curve relating DNA fragment size (molecular weight markers) and distance migrated (mm). 58

Fig. 3.10. (A) Agarose gel electrophoresis of the products of the reaction used to amplify CA genomic DNA including both exon and intron gene components. (B) The sequences of the DNA amplified, specifically the first exon and first intron amplified. 60

Fig. 3.11. Matrix alignment plot showing (A) similarity between the AZM4_68974 assembly and the previously sequenced 5′-end of the maize CA2 gene, and (B) similarity between the AZM4_68974 assembly and the genomic library clone. Similarity is based on a minimum alignment of 80% in a 20 bp window (hash value 6). 62

Fig. 3.12. Schematic representation of the AZM4_68974 assembly, with exons indicated that corresponded to the CA2 cDNA sequence. 63

Fig. 3.13. Matrix alignment plot showing (A) similarity between the AZM4_68973 assembly and the CA2 gene sequence, and (B) similarity between the AZM4_68973 assembly and the genomic DNA library clone. Similarity is based on a minimum alignment of 80% in a 20 bp window (hash value 6). 64

Fig. 3.14. Schematic representation of the AZM4_68973 genomic sequence, with exons indicated that corresponded to the CA2 cDNA sequence. 65

Fig. 3.15. Matrix alignment plot showing similarity between the AZM4_23203 assembly and the AY109272 mRNA sequence, indicating the exon/intron structure. Similarity is based on a minimum alignment of 80% in a 20 bp window (hash value 6). 66

Fig. 3.16. Schematic representation of the AZM2_23203 genomic sequence. 66

Fig. 3.17. ClustalW amino acid alignment of the AY109272 and DQ246083 translated sequences with the maize CA amino acid sequence of Repeat A. 67

Chapter 4

Fig. 4.1. Structural differences between monocot and dicot plants (http://biology.clc.uc.edu/graphics/bio106/bean.jpg). 78

Fig 4.2. ClustalW nucleotide alignment of Repeat A, Repeat B and Repeat C of the maize CA cDNA sequences. 81

Fig. 4.3. Schematic representation of the primer binding sites on the CA cDNA sequence in the three reactions used to identify the transcription initiation site using a maize cDNA library as template. Figure not to scale. 85

Fig. 4.4. Schematic representation of the regions of the CA transcript amplified. Ldr: amplification of the 5′-leader sequence, 276 bp Ins: amplification of a region of the unique 276 bp insert of CA2, Repeat: amplification within the repeating region of the CA transcript. Due to the repetitive sequence of the repeating regions, the primers used to amplify the repeat product would not distinguish between Repeat A, Repeat B or Repeat C. The CA2 cDNA schematic was taken from Burnell and Ludwig, (1997). 87

Fig. 4.5. Phylogeny of the beta-CAs showing separation of monocot and dicot species. Repeat A, Repeat B and Repeat C refer to the maize CA amino acid sequences. 89

Fig. 4.6. Schematic representation of (A) the maize CA2 gene and (B) the rice CA gene. Introns and exons are drawn to scale and indicated as lines or boxes respectively. 91

Fig. 4.7. Analysis of gene structure for (A) the maize CA2 gene, and (B) the rice CA gene. The length of the introns and exons is compared, represented by the histograms, while the line graph shows the percentage of G/C residues composing these regions. 92

Fig. 4.8. Location of CpG islands in the 5′-upstream region of the maize and rice CA gene sequences. 93

Fig. 4.9. Agarose (1.5%) gel electrophoresis of the products obtained from PCR on a maize cDNA library. Lane 1: CA1R and T3 primers; 2: InsR and T3 primers; 3: CA7 and T3 primers; and 4: negative control (minus template). 95

Fig. 4.10. Comparison of the maize and rice genomic DNA sequences representing 1,000 bp upstream of the translation initiation codon, ATG, of the CA gene for both species. Similarity is based on a minimum similarity of 50% in a 20 bp window (hash value 6). 100

Fig 4.11. ClustalW alignment of the deduced amino acid sequence of Repeat A of the maize CA2 cDNA sequence and the rice chloroplastic CA (not including the chloroplast transit peptide). 100

Fig. 4.12. The relative transcript abundance as determined by semi-quantitative RT-PCR (Fig. 4.13-4.18) using maize roots, and leaf base, middle and tip as sources of template RNA. 101

Chapter 5

Fig. 5.1. Schematic representation of the three beta-CA cDNAs from maize. The repeat sequences are approximately 600 bp and are labelled A, B and C. The black shaded box represents the 276 bp insert unique to CA2. The dark grey shaded boxes represent the 5′-leader sequences and 3′-untranslated regions. 120

Fig. 5.2. Alignment of the amino acid sequences of Repeat A (RA), Repeat B (RB) and Repeat C (RC). 121

Fig. 5.3. Reaction catalyzed by CA. 122

Fig. 5.4. The two stage ping pong reaction mechanism described for alpha-CAs (Kimber and Pai, 2000). E represents the enzyme and B represents the buffer molecule. 123

Fig. 5.5. Schematic representation of the components of the CA transcripts that were expressed and used to analyse CA activity. The histidine tag is represented as a black box and the 5′-leader sequence is represented as a putative transit peptide. 125

Fig 5.6. Schematic representation of Repeat A showing the *Nco*I recognition site within the Repeat A sequence. 127

Fig. 5.7. Western blots and Coomassie blue-stained SDS-PAGE of CA1, indicated with an arrow. 5 µl of 2.5 mg.ml^{-1} protein sample was loaded per lane (12.5 µg total protein). For each sample, 5 µl of the 200 mM imidazole eluted fraction is represented, which was then used for CA activity analysis. α-pCA refers to α-CA antibodies that have been pre-incubated with a crude extract of bacterial proteins for 3 h at 4ºC (Section 5.2.4.3). 134

Fig. 5.8. Coomassie blue-stained SDS-PAGE and Western blots of Repeat A, indicated with an arrow. 5 µl of 1.5 mg.ml^{-1} protein sample was loaded per lane (7.5 µg total protein). For each sample, 5 µl of the 200 mM imidazole eluted fraction is represented, and which was then used for CA activity analysis. 135

Fig. 5.9. The amino acid sequence of Repeat B, with the additional 16 amino acids present at the end of the sequence shown in bold. The stop codon is represented by an asterisk. 135

Fig. 5.10. Coomassie blue-stained SDS-PAGE and Western blots of Repeat B, indicated with an arrow. 5 µl of 5.4 mg.ml^{-1} protein sample was loaded per lane (27 µg total protein). For each sample, 5 µl of the 200 mM imidazole eluted fraction is represented, and which was then used for CA activity analysis. α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2). 136

Fig. 5.11. Coomassie blue-stained SDS-PAGE and Western blots of Repeat C, indicated with an arrow. 5 µl of protein was loaded per lane, equivalent to 10.5 µg of total protein loaded (2.1 mg.mL^{-1}). For each sample, 5 µl of the 200 mM imidazole eluted fraction is represented, and which was then used for CA activity analysis. α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2). 136

Fig. 5.12. Coomassie blue-stained SDS-PAGE and Western blots of CA4, with the predicted molecular weight, 52.8 kDa, and the observed immuno-reactive band at 28 kDa indicated with arrows. 5 µl of the 200 mM imidazole eluted protein fraction was loaded in each lane, equivalent to 10.5 µg of total protein loaded (2.1 mg.mL^{-1}). α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2). 137

Fig. 5.13. Coomassie blue-stained SDS-PAGE and Western blots of CA5, indicated with an arrow just below 28 kDa. 5 µl of the 200 mM imidazole eluted protein fraction was analyzed, the equivalent of 9.75 µg protein (concentration was 1.95 mg.ml^{-1}). α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2). 138

Fig. 5.14. Coomassie blue-stained SDS-PAGE and Western blots of CA6, indicated with an arrow. 5 µl of the 200 mM imidazole eluted protein fraction was analyzed, the equivalent of 10.5 µg protein (concentration was 2.1 mg.ml^{-1}). α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2). 139

Fig. 5.15. Coomassie blue-stained SDS-PAGE and Western blots of CA7. The protein subunit is indicated with an arrow. 5 µl of the 200 mM imidazole eluted protein fraction was analyzed, the equivalent of 12.5 µg protein (concentration was 2.5 mg.ml^{-1}). α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2). 139

Fig. 5.16. Western blots of the bacterial extraction of proteins without recombinant CA (-), and (+) containing CA7. Non-specific hybridization is indicated with an arrow at 55 kDa and the protein subunit is indicated with an arrow at 28 kDa. α-pCA refers to α-CA antibodies that have been pre-incubated with a crude extract of bacterial proteins for 3 h at 4ºC (Section 5.2.4.3). 140

Fig. 5.17. Coomassie blue-stained SDS-PAGE and Western blots of CA8. The protein subunit is indicated with an arrow. 5 µl of the 200 mM imidazole eluted protein fraction was analyzed, the equivalent of 15.95 µg protein (concentration was 3.2 $mg.ml^{-1}$). α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2). 141

Fig. 5.18. The activity of the bacterially expressed CA enzymes expressed as $units.mg^{-1}$. One unit of activity is calculated as µmol of carbon dioxide consumed per second. RA is Repeat A and RB is Repeat B. 142

Fig. 5.19. The activity of CA1 (expressed as units) in the presence of 1 mM acetazolamide, 1 mM 6-ethoxyzolamide, 10 mM 1,10-phenanthroline or 1% diethylpyrocarbonate. The measurements taken are compared to the control reactions, which were performed in the presence of 10% methanol (acetazolamide, 6-ethoxyzolamide and 1,10-phenanthroline) or 10% ethanol (diethylpyrocarbonate). *p= <0.05, ** p=<0.01, *** p=<0.001 Mann-Whitney U Test. 144

Fig. 5.20. The activity of Repeat A (expressed as units) in the presence of 1 mM acetazolamide, 1 mM 6-ethoxyzolamide, 10 mM 1,10-phenanthroline or 1% diethylpyrocarbonate. The measurements taken are compared to the control reactions, which were performed in the presence of 10% methanol (acetazolamide, 6-ethoxyzolamide and 1,10-phenanthroline) or 10% ethanol (diethylpyrocarbonate). *p= <0.05, ** p=<0.01, *** p=<0.001 Mann-Whitney U Test. 144

Fig. 5.21. The activity of Repeat B (expressed as units) in the presence of 1 mM acetazolamide, 1 mM 6-ethoxyzolamide, 10 mM 1,10-phenanthroline or 1% diethylpyrocarbonate. The measurements taken are compared to the control reactions, which were performed in the presence of 10% methanol (acetazolamide, 6-ethoxyzolamide and 1,10-phenanthroline) or 10% ethanol (diethylpyrocarbonate). *p= <0.05, ** p=<0.01, *** p=<0.001 Mann-Whitney U Test. 145

Fig. 5.22. The activity of CA4 (expressed as units) in the presence of 1 mM acetazolamide, 1 mM 6-ethoxyzolamide, 10 mM 1,10-phenanthroline or 1% diethylpyrocarbonate. The measurements taken are compared to the control reactions, which were performed in the presence of 10% methanol (acetazolamide, 6-ethoxyzolamide and 1,10-phenanthroline) or 10% ethanol (diethylpyrocarbonate). *p= <0.05, ** p=<0.01, *** p=<0.001 Mann-Whitney U Test. 145

Fig. 5.23. The activity of CA5 (expressed as units) in the presence of 1 mM acetazolamide, 1 mM 6-ethoxyzolamide, 10 mM 1,10-phenanthroline or 1% diethylpyrocarbonate. The measurements taken are compared to the control reactions, which were performed in the presence of 10% methanol (acetazolamide, 6-ethoxyzolamide and 1,10-phenanthroline) or 10% ethanol (diethylpyrocarbonate). *p= <0.05, ** p=<0.01, *** p=<0.001 Mann-Whitney U Test. 146

Fig 5.24. ClustalW alignment of the deduced amino acid sequences of Repeat A of the maize CA2 cDNA sequence and the rice chloroplastic CA (not including the chloroplast transit peptide). 148

Fig. 5.25. Alignment of the active, expressed CA amino acid sequences (RB – Repeat B, CA5 – Repeat C/A chimera, and RA – Repeat A) with the inactive amino acid sequences (RC – Repeat C and CA7 – Repeat A/C chimera). Differences in amino acid residues for both groups of sequences are indicated with bolding. Asterisks indicate potential zinc ligands, and filled circles indicate potential zinc ligands that when mutated, resulted in CA that bound zinc poorly (Bracey *et al*., 1994; Cronk *et al*., 2001). The open circle indicates the glutamine residue suggested to act as a hydrogen bond donor in the Arabidopsis beta-CA (Rowlett *et al*., 1994). The triangle indicates a cysteine residue that is conserved between monocot and dicot plant sequences, which in pea is essential for structure and enzyme activity (Björkbacka *et al*., 1997). The predicted secondary structure was different for the active and inactive sequences in the region indicated by the grey line. 151

Chapter 6

Fig. 6.1. Ribbon diagram of the crystal structure of the pea CA monomer, from N-terminus (blue) to C-terminus (red), with the zinc atom in the active site indicated by a blue sphere (Kimber and Pai, 2000). 163

Fig. 6.2. Ribbon diagram of the crystal structure of the *E. coli* CA monomer (Cronk *et al.*, 2001). Helices are shown in red, and beta-strands in yellow. The magenta sphere indicates the location of the zinc atom. 163

Fig. 6.3. Ribbon diagram of the crystal structure of the *P. purpureum* CA (Mitsuhashi *et al*., 2000). The N-terminal is marked and coloured blue, the C-terminal is also marked and coloured green. The zinc atoms are shown as red spheres. 164

Fig. 6.4. Schematic representation of primer annealing sites relative to the 276 bp insert of the CA2 cDNA sequence (gi:606810). The 5′-leader corresponds to the first exon, while the third exon encodes the start of Repeat A. Figure not to scale. 168

Fig. 6.5. Probability of the translated sequence of Exon 1 (N-terminal sequence) acting as a chloroplast transit peptide. 174

Fig. 6.6. Alignment of the rice and maize CA chloroplast transit peptide amino acid sequences. The putative cleavage site of the chloroplastic transit peptide is indicated with an arrow. 175

Fig. 6.7. ClustalW alignment of Repeat A of the maize CA sequence and the rice chloroplastic CA (not including the chloroplast transit peptide). 175

Fig. 6.8. Hydrophobicity plot based on the amino acid sequence of Repeat B (Appendix 6.2). 176

Fig. 6.9. Graphical output from the DAS trans-membrane prediction server. The full line represents a strict cutoff, while the dotted line represents a loose cutoff. 177

Fig. 6.10. Secondary structure predictions based on the amino acid sequence of Repeat B. Alpha-helices are indicated (H), and beta-sheets (b). The prediction based on use of the PHD server is indicated by 'PHD' and the Softberry secondary structure prediction by 'Sber'. These are shown in red where both algorithms have predicted the same secondary structure. 178

Fig. 6.11. Coomassie blue-stained gel and corresponding Western blot of the Repeat B sample used for circular dichroism analysis. 5 µl of protein was loaded per lane, equivalent to 7.5 µg of total protein (1.5 $mg.ml^{-1}$ protein concentration). 179

Fig. 6.12. CD spectra for Repeat B 179

Fig. 6.13. Coomassie blue-stained discontinuous native PAGE of Repeat B (7.5 µg total protein). 180

Fig. 6.14. PCR products generated from the primer combinations indicated in Table 6.1. The DNA fragments shown are representative of a number of repetitions of these reactions. 181

Fig. 6.15. Alignment of the genomic sequence (AZM4_68974) with the CA2 sequence (as reported, accession gi:606810) at the start of the second exon. The reported (CA2 cDNA) and the observed reading frames are presented, with the stop codon indicated with an asterisk. 181

Fig. 6.16. Alignment of the reported CA2 sequence (gi:606810) with the sequence obtained from analysis of maize leaf mRNA. Numbering is from the start of the second exon representing the 276 bp insert unique to CA2. 182

Fig. 6.17. Western blot of 10 µl of the preparation of maize leaf proteins (1.3 $mg.ml^{-1}$) using α-CA antibodies. 183

Fig. 6.18. Typical elution profile of the proteins used for column calibration. The absorbance at 280 nm was used to determine in which fractions ferritin and cytochrome c (Cyto c) had eluted, while the rate of NADH oxidation measured at 340 nm was used to determine in which fractions LDH and MDH had eluted. 184

Fig. 6.19. Typical standard curve relating native molecular weight (M_r) to the distribution coefficient (K_{dist}) for four known protein samples, with the equation to the line indicated. 185

Fig. 6.20. Detection of CA in fractions eluted from the gel filtration column by ELISA. 185

Fig. 6.21. Western blot showing a 52 kDa immuno-reactive band in Fraction 7 after concentration by TCA precipitation. 187

Fig. 6.22. Analysis of CA by Western blotting after precipitation with ammonium sulphate (65% pellet), and analysis of eluted fractions (0 M, 0.2 M, 0.5 M and 1 M NaCl) obtained after subsequent separation by ion exchange chromatography in the presence of Triton X-100. 188

Fig. 6.23. Alignment of the rice, maize and *F.bidentis* CA putative transit peptide sequences. The cleavage site of the chloroplast transit peptide is indicated with an arrow. 190

Abbreviations

× *g*	times gravity
°C	degrees Celsius
A_{260}	absorbance at 260 nm
aa	amino acid
ABA	abscisic acid
ABRE	ABA-responsive element
ARE	anaerobic responsive element
ATP	adenosine-5′-triphosphate
BLAST	basic local alignment search tool
bp	base pair
BSA	bovine serum albumin
CA	carbonic anhydrase
CA-RP	carbonic anhydrase-related protein
CCM	carbon concentrating mechanism
CD	circular dichroism
cDNA	complementary deoxyribonucleic acid
cm	centimetre
dCTP	deoxycytidine triphosphate
DEAE	diethylaminoethyl
dicot	dicotyledon
DRE	dehydration-responsive element
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
Dof	DNA-binding with one finger
dpm	disintegrations per minute
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
ELISA	enzyme linked immuno-sorbent assay
EREBP	ethylene-responsive element binding protein
EXAFS	extended X-ray absorption fine structure
Fig.	figure
g	gram
GST	glutathione *S*-transferase
h	hour
HRP	horse radish peroxidase
IgG	Immunoglobulin G
IPTG	isopropyl-β-D-thiogalactopyranoside
IMAC	immobilized metal affinity chromatography
kb	kilo base pair
k_{cat}	catalytic rate of an enzyme
K_d	dissociation constant
K_{dist}	distribution coefficient
kDa	kilodaltons
K_m	Michaelis-Menten constant
L	litre
LB	Luria Broth

LDH	lactate dehydrogenase
M	molar
MDH	malate dehydrogenase
mg	milligram
min	minute
ml	millilitre
mm	millimetre
mM	millimolar
monocot	monocotyledon
mRNA	messenger ribonucleic acid
Mw	molecular weight
n	number
NAD	nicotinamide adenine dinucleotide
NCBI	National Center for Biotechnology Information
NIP	nearly identical paralog
ng	nanogram
nm	nanometre
ocs	octopine synthase
OD	optical density
ORF	open reading frame
p	probability
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PEP	phospho*enol*pyruvate
PEP-CK	PEP carboxykinase
pfu	plaque forming unit
pH	$-\log_{10}[H^+]$
pI	isoelectric point
pmol	picomole
PPDK	pyruvate orthophosphate dikinase
PR gene	pathogenesis-related gene
PS II	Photosystem II
PVDF	polyvinylidene fluoride
RA	Repeat A
RACE	rapid amplification of cDNA ends
RB	Repeat B
RC	Repeat C
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
rpm	revolutions per minute
rRNA	ribosomal RNA
RT-PCR	reverse-transcriptase PCR
Rubisco	ribulose-1,5-bisphosphate carboxylase/oxygenase
RuBP	ribulose-1,5-bisphosphate
s	second
SA	salicylic acid
SABP3	salicylic acid-binding protein 3
SAP	shrimp alkaline phosphatase
SDS	sodium dodecyl sulphate

snRNPs	small nuclear ribonucleoproteins
TCA	trichloroacetic acid
TE	10 mM Tris-HCl pH 7.5, 1 mM EDTA
Tris	tris (hydroxymethyl) aminomethane
μg	microgram
μl	microliter
μm	micrometer
μM	micromolar
μmol	micromole
UTR	untranslated region
UV	ultra violet light
V	volts
v_e	volume of the elution peak height
v_o	void volume
V_{max}	maximum reaction rate
vol (or) v	volume
WOC	water-oxidizing complex
w	weight
X-Gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside

Publications

Tems, U. and Burnell, J. "Carbonic Anhydrase Isozymes from *Zea mays*." Poster. Combio 2006 Combined Conference (ASBMB, ASPS, AuPS, ANZSCDB, NZSBMB and NZSPP), Brisbane, Australia.

Tems, U. and Burnell, J. (2009) "The structure of the maize β-carbonic anhydrase gene contains two repeat regions with expression of a single repeat region producing an active enzyme." *(Manuscript in preparation)*.

CHAPTER 1

1. Introduction

Carbonic anhydrase (CA; E.C. 4.2.1.1) is a ubiquitous enzyme that participates in a range of biological functions including carboxylation reactions, acid-base balance, ion exchange and involvement in the cellular processes of respiration and photosynthesis, in a wide range of organisms (Tashian, 1992). While the structure and function of CA has been resolved in many species of animals and algae, the sub-cellular location and role of CA in higher plants, particularly those that use the C_4 photosynthetic pathway, requires further investigation.

The C_4 photosynthetic pathway is a mechanism whereby many species of plants eliminate energetically wasteful photorespiration and concentrate carbon for use in the photosynthetic carbon reduction cycle. CA catalyzes the first reaction in the C_4 photosynthetic pathway, hydrating carbon dioxide that has diffused into the mesophyll cell cytoplasm to bicarbonate, providing an inorganic carbon source for the C_4 pathway (Hatch and Burnell, 1990).

CA gene families

The genes encoding CA are found in species from all three phylogenetic domains of life and represent a diverse multi-gene family. A lack of primary sequence similarity between the gene families, despite conservation of enzymatic mechanism and in some cases secondary structure, provides an example of convergent evolution of catalytic function. Nearly all CAs are zinc metallo-enzymes that share a similar mechanism for the interconversion of carbon dioxide and bicarbonate (Tripp *et al.,* 2001; Hewett-Emmett and Tashian, 1996).

There are three main CA gene families classified according to primary sequence, which have been designated alpha-, beta- and gamma-CAs. The first CA isolated belonged to the alpha-CA gene family, and is involved in carbon dioxide diffusion in red blood cells in mammals (Meldrum and Roughton, 1933). CAs from the beta- and gamma-gene families are most highly represented in bacteria and higher plants. The fourth CA gene family was first discovered in the marine diatom *Thalassiosira weissflogii*, and classified as a delta-CA due to lack of sequence similarity with either alpha-, beta- or gamma-CAs. This species also contains a unique cadmium-dependent CA, which has been assigned to the epsilon-CA gene family (Lane *et al.,* 2005; McGinn and Morel, 2008).

Alpha CA gene family

Alpha-CAs are the most extensively studied due to identification of the enzyme in humans and its association with a wide range of physiological as well as pathological processes. Human CA isozymes are present in blood, muscle tissue, mitochondria and saliva (Table 1.1). Tissue-specific expression, unique physiological functions and differing inhibitor sensitivity have aided purification and characterization of these isozymes. For example, CA I and CA II can be separated by affinity chromatography based on differences in sulphonamide and monovalent anion binding of the two isozymes (Bergenhem, 1996). Antigenic differences can be used to distinguish the alpha-, beta- and gamma-CAs, supported by analysis of gene sequences. Antigenic differences also exist within the alpha-CA gene family, dividing the gene family into two groups. The first group consists of the cytoplasmic and mitochondrial CAs, while the second group includes the membrane-bound and secreted

CAs, such that antibodies raised against CA VI have no cross-reactivity against CA II (Fernley, 1988; Jiang and Gupta, 1999).

Table 1.1. Classification of the human CA isozymes (Fernley, 1998; Tashian, 1992; Lehtonen *et al.,* 2004; Vullo *et al.,* 2005; Di Fiore *et al.,* 2008; Hilvo *et al.,* 2008; Bergenhem 1996).

Isozyme	**Cellular Location**	**Tissue Distribution**	**Molecular Weight (kDa)**	**Activity**
I	Cytoplasm	Erythrocytes	30	Moderate
II	Cytoplasm	Widespread	30	High
III	Cytoplasm	Red skeletal muscle, liver	30	Low
IV	Cell membrane	Lung, kidney	52 68	High
V	Mitochondria	Liver, kidney	29	High
VI	Secreted (extracellular)	Parotid salivary gland	45	High
VII	Cytoplasm	Salivary glands, brain	-	High
VIII	Cytoplasm			CA-RP
IX	Cell membrane	Gastric mucosa, tumour-associated	50	High
X	Cytoplasm	-	-	CA-RP
XI	Cytoplasm	-	-	CA-RP
XII	Cell membrane	Eye, lung, kidney, tumour-associated	-	Yes
XIII	Cytoplasm	Reproductive tract, thymus, colon	30	Low
XIV	Cell membrane	-	-	Yes
XV	Cell membrane	Membrane associated	29	Yes

The most recent alpha-CA identified, CA XV, is encoded by a pseudo-gene in humans and chimpanzees, and there are no mRNA sequences in expressed sequence tag databases for this CA. In rodents, birds and fish an active form exists, which plays a role in kidney function and pH regulation (Hilvo *et al.*, 2005). Several CA-related proteins (CA-RPs) share sequence similarity to the alpha-CA gene family but show no CA activity and have unknown biological functions (Table 1.1; Sly and Hu, 1995; Elleby *et al.,* 2000). The

CA-RPs do not have catalytic activity due to a change in the amino acid structure at the active site, however they have a similar secondary structure (Vullo *et al.,* 2005). Only two point mutations were required to alter a murine CA-RP to an active form. These mutations involved a change from an arginine to a histidine at position 117, and a glutamate to a glutamine at position 115, residues that coordinate the active site zinc ion (Sjöblom *et al.,* 1996).

The reaction catalyzed by CA is the reversible hydration of carbon dioxide to bicarbonate, which can also occur spontaneously. CA II has one of the fastest reaction rates known for any enzyme with a k_{cat} of 1.4 x 10^6s^{-1} (Silverman, 1991; Tashian, 1992). CA IX has recently been characterized and has a similar k_{cat} as CA II. It contains a proteoglycan-like domain along with a CA domain, and the expressed recombinant protein containing the proteoglycan-like domain had an unexplained faster reaction rate than expression of only the CA domain (Hilvo *et al.,* 2008). CA VII is the second most active cytosolic isozyme, and has approximately 70% of human CA II activity (Vullo *et al.,* 2005). While the reaction catalyzed by CA II can occur spontaneously, defects in this isozyme are responsible for causing osteopetrosis (a rare congenital disorder in which the bones become overly dense), renal tubular acidosis and brain calcification in humans (Sly and Hu, 1995). These defects have clarified the function of CA II. In contrast, while CA I is highly abundant, no abnormalities are obvious when this isozyme is absent or mutagenically inactivated (Tashian, 1992).

Like other alpha-CAs, the secondary structure of CA II is dominated by beta-sheets with only a single active-site domain (Aronsson *et al.,* 1995; Mitsuhashi *et al.,* 2000). The active-site structure of alpha-CAs has three conserved histidine residues coordinated with a catalytic zinc ion, which participates in the reaction by interacting with the substrate molecule carbon dioxide (McCall *et al.,* 2000). The proposed reaction mechanism based on experimental studies of human CA II involves a zinc hydroxide for carbon dioxide hydration (Fig 1.1).

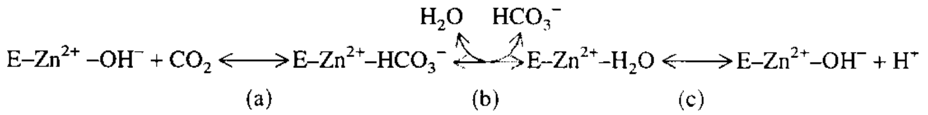

H2O HCO3−
E–Zn2+ –OH− + CO2 ⟷ E–Zn2+–HCO3− ⟷ E–Zn2+–H2O ⟷ E–Zn2+–OH− + H+
(a)
(b)
(c)

Fig 1.1. *Reaction mechanism of CA showing (a) nucleophilic attack of the metal-activated hydroxide ion on carbon dioxide, (b) ligand exchange of the product bicarbonate for a water molecule, and (c) regeneration of the zinc hydroxide form of the enzyme (Cronk et al., 2001).*

Alpha-CAs have been identified and characterized in animals, insects, algae, the nematode *Caenorhabditis elegans,* bacteria such as *Helicobacter pylori,* and other prokaryotes. An alpha-CA gene in *C. elegans* encodes two CA isozymes, the expression of which contributes to the ability of the nematode to survive in the pH range of 3 to 10 (Hall *et al.,* 2008). In coral species, alpha-CA plays a role in biomineralization, producing bicarbonate for the generation of calcium carbonate in hard tissue (Moya *et al.,* 2008). In the giant clam *Tridacna gigas*, a 70 kDa alpha-CA that contained two active sites with two CA domains was identified, and this characteristic was unique from other animal CAs of the alpha-gene family (Leggat *et al.,* 2005). An alpha-CA has been identified in the pathogenic bacteria *Neisseria gonorrhoeae,* which has a similar reaction mechanism and secondary structure to human CA II, however it exhibits only 35% amino acid sequence similarity (Elleby *et al.,* 2000). In other pathogenic species such as the mosquito *Aedes aegypti,* comparison of alpha-CA structure and function has enabled the development of specific inhibitors (Fisher *et al.,* 2006).

The alpha-CA gene family has been characterized in the model higher plant Arabidopsis (*Arabidopsis thaliana*). Sequencing of the Arabidopsis genome led to the discovery of eight alpha-CA isozymes, as well as six beta-CAs and five gamma-CAs (Parisi *et al.,* 2004; Fabre *et al.,* 2007). Of the eight alpha-CAs, transcripts were only identified for the first three isozymes using a publicly available expressed sequence tag library, and the remaining five isozymes could not be amplified by reverse transcriptase-PCR. However screening of a cDNA library did isolate CA7, while CA4 was associated with the thylakoid membrane (Friso *et al.,* 2004).

Initial reports of an alpha-CA in the eukaryotic algae *Chlamydomonas reinhardtii* associated with the thylakoid membrane were based on similarities in the inhibition of CA activity and photosystem II (PS II) electron transport by anions and acetazolamide (Moubarak-Milad and Stemler, 1994). A 29.5 kDa intracellular alpha-CA in *C. reinhardtii* immuno-localized in association with the chloroplast thylakoid membrane (Karlsson *et al.,* 1995; Karlsson *et al.,* 1998). A thylakoid-associated alpha-CA in *Pisum sativum* (pea), distinct from the soluble cytosolic form of the enzyme, was inhibited by diuron and hydroxylamine, both of which inhibit PS II (Stemler, 1997). The function of the thylakoid-associated alpha-CA was proposed to be the generation of carbon dioxide in the acidic environment of the thylakoid lumen, using bicarbonate that had been pumped into the lumen from the cytoplasm, rather than being directly linked to the function of PS II (Park *et al.,* 1999). Energy required to pump bicarbonate into the thylakoid lumen would be generated by the pH gradient created by photosynthetic electron transport (Fig. 1.2). Neither electron transport nor adenosine triphosphate (ATP) were limited in a *C. reinhardtii* mutant strain that lacked the thylakoid-associated CA, supporting a role for this CA in carbon metabolism rather than PSII function (Hanson *et al.,* 2003).

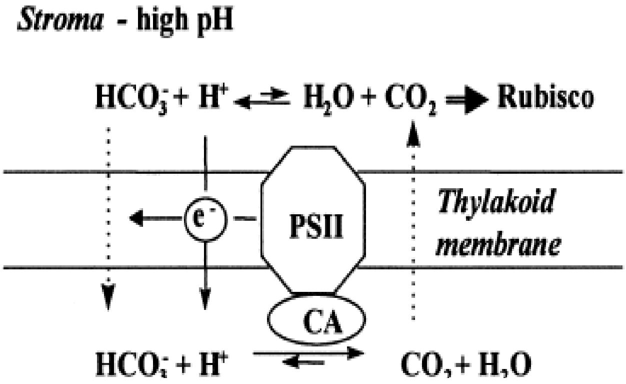

Stroma - high pH
$HCO_3^- + H^+ \rightleftharpoons H_2O + CO_2 \rightarrow$ Rubisco
e-
PSII
Thylakoid membrane
CA
$HCO_3^- + H^+ \rightleftharpoons CO_2 + H_2O$

Fig. 1.2. *Diagrammatic representation of the role of CA in the thylakoid lumen, supplying carbon dioxide to Rubisco as part of the carbon concentrating mechanism in* C. reinhardtii *(Park* et al.*, 1999).*

However, several studies have confirmed that thylakoid-associated alpha-CAs are involved directly in PS II function (Moubarak-Milad and Stemler, 1994; Lu and Stemler, 2002). The strain of *C. reinhardtii* lacking the thylakoid-associated CA was unable to oxidize water at PS II (Villarejo *et al.,* 2002). The process of water oxidation involves the water-oxidizing complex (WOC), a manganese cluster and PS II-associated proteins in the thylakoid lumen. In the absence of the thylakoid-associated CA, the manganese cluster became unstable and electron donation was impaired. In *Zea mays* (maize), a species which lacks at least a beta-CA isoform in mesophyll cell chloroplasts, a 33 kDa CA isozyme was identified in the thylakoids and confirmed to be an alpha-CA due to cross-reactivity with antibodies generated against the alpha-CA from *C. reinhardtii.* Furthermore, these studies demonstrated a second CA isozyme present in the maize thylakoid membrane, the activity of which was still detectable after a calcium chloride wash, while in pea thylakoids up to four CAs have been identified (Lu and Stemler, 2002; Rudenko *et al.,* 2007). In maize, the absence of ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) in mesophyll chloroplasts supports the hypothesis that thylakoid-associated CA functions with PS II.

Beta CA gene family

The beta-CAs are present in higher plants, algae, eubacteria and archaea (Moroney *et al.,* 2001). Kinetic studies of beta-CAs from higher plants have shown them to be catalytically as efficient as human CA II (Mitsuhashi *et al.,* 2000). While beta-CA genes share significant sequence homology, there is considerable variation in subunit size, native molecular weight and physiological function within this group (Table 1.2). The beta-CAs found in prokaryotes exhibit lower sequence similarity than the plant isozymes (Hiltonen *et al.,* 1998).

Table 1.2. Characteristics of the beta-CA gene family (Eriksson *et al.,* 1998; Giordano *et al.,* 2003; Mitra *et al.,* 2004; Mitsuhashi *et al.,* 2000; Hiltonen *et al.,* 1998; Satoh *et al.,* 2001; Cronk *et al.,* 2001; Kusian *et al.,* 2002; Yang *et al.,* 1985; Graham *et al.,* 1984; Atkins *et al.,* 1972a; Fabre *et al.,* 2007).

Organism	Number of Isozymes	Subunit size (kDa)	Suggested Function
Prokaryotes and Algae			
Chlamydomonas reinhardtii	6	28.5 27 35	Nitrogen metabolism, CCM
Porphyridium purpureum	1	55	CCM
Coccomyxa	1	25	Not known
Phaeodatylum tricornutum	1	28	CCM
Escherichia coli	2	24	Cell metabolism, pH regulation
Ralstonia eutropha	1	25	CCM
Higher Plants			
Pisum sativum (Pea)	2	24.2 30	Photosynthesis
Spinacia oleracea (Spinach)	2	35 30	Photosynthesis
Nicotiana tabacum (Tobacco)		30 24	Hypersensitive defense response
Gossypium hirsutum (Cotton)	2	35 24	Post-germinative growth, Photosynthesis
Amaranthus cruentus	2		Photosynthesis
Sorghum bicolor (Sorghum)	3		Photosynthesis
Zea mays (Maize)	3	52 45 28	Photosynthesis
Flaveria sp.	3	35 31	Photosynthesis

Two nearly identical beta-CA isozymes have been identified in the green algae *C. reinhardtii,* a species that also contains alpha-CA enzymes. A 28.5 kDa chloroplastic beta-CA was discovered by immunological analysis and suggested to have a role in photosynthesis, particularly in the carbon concentrating mechanism (CCM; Mitra *et al.,* 2004). Supporting this, expression of CA in *C. reinhardtii* is dependent on carbon dioxide concentration, with low carbon dioxide levels inducing high levels of CA mRNA (Eriksson *et al.,* 1998). A beta-CA has also been localized to the mitochondria, where it provides bicarbonate ions for the synthesis of carbon skeletons for nitrogen assimilation (Giordano *et al.,* 2003). The green alga *Coccomyxa* contains many species that do not have a CCM, but have high intracellular CA activity. This activity was located in the cytoplasm rather than the chloroplast (Hiltonen *et al.,* 1998).

A 55 kDa beta-CA has been purified from the unicellular red alga *Porphyridium purpureum* that has a role in the CCM by providing adequate substrate carbon dioxide levels (Mitsuhashi and Miyachi, 1996). This enzyme may be a product of gene duplication, with the N- and C-terminal halves of the mature enzyme having sequence homology as well as two zinc-containing active sites (Mitsuhashi *et al.,* 2000). This CA associates as a dimer, with a native molecular weight of 110 kDa and has four zinc-containing active sites. Another example of this is the alpha-CA gene from *Dunaliella salina,* which encodes a 63 kDa protein that appears to have two active sites (Moroney *et al.,* 2001). Also the alpha-CA from *T. gigas* has two CA domains and contains two active sites (Leggat *et al.,* 2005). It has been suggested that this is a similar arrangement that arises from the association of subunits of other beta-CAs producing active enzyme (Mitsuhashi *et al.,* 2000).

An intracellular beta-CA has been characterized from the marine diatom *Phaeodactylum tricornutum* which has a subunit molecular weight of 28 kDa (Satoh *et al.,* 2001). The expression of this CA is dependent on low carbon dioxide concentrations, implying a role in the CCM of this organism. The only other CA that has been purified from a marine diatom, *T. weissflogii,* showed no sequence homology to any of the CA gene families and represents a new delta-CA gene family (McGinn and Morel, 2008).

Two beta-CA isozymes are present in *Escherichia coli,* enabling cyanate to be used as a nitrogen source by replenishing bicarbonate levels when cyanate concentrations are high (Cronk *et al.,* 2001). The activity of the second CA is pH-dependent, suggesting that the enzyme may be subject to pH-mediated regulation in the cell. A beta-CA gene is also present in *Ralstonia eutropha,* a chemoautotroph (Kusian *et al.,* 2002). This gene was discovered in a mutant strain in which growth was inhibited under low carbon dioxide concentrations. The sequence homology between this gene and the beta-CAs from *E. coli* and *P. purpureum* confirmed that it belongs to the beta-CA family.

CA is usually a zinc metallo-enzyme, with a zinc ion required for CA activity and located in the active site of the enzyme subunit. In the dicot *Phaseolus vulgaris* there was a correlation between CA enzyme activity and zinc deficiency, while other enzymes including Rubisco, glycolate oxidase and malic enzyme were not affected (Edwards and Mohamed, 1973). Unlike the alpha-CAs, where coordination of zinc in the active site is achieved by the imidazole groups of three histidine residues, the coordination of the zinc molecule in beta-CAs is reliant on only one histidine and two cysteine residues (Cronk *et al.,* 2001). Previously, alignments of beta-CA amino acid sequences had enabled these residues to be predicted based on residue conservation (Hewett-Emmett and Tashian, 1996). The histidine residue is essential for a proton-shuttling function. Substituting histidine with alanine at this position altered the kinetic values for the enzyme, and only imidazole-type buffers were able to replace the proton-shuttling function of the histidine residue (Björkbacka *et al.,* 1999).

A phylogenetic analysis of the beta-CA family shows three distinct monophyletic groups representing the eubacteria, monocotyledon (monocot) plants and dicotyledon (dicot) plants (Fig. 1.3; Hewett-Emmett and Tashian, 1996). Flowering plants are classed as monocots or dicots depending on the number of cotyledons, or seed leaves, found in the embryo. There is greater CA sequence homology within monocot and dicot classes than between them, although the sequence homology between monocot and dicots is still significant (Fig. 1.4; Moroney *et al.,* 2001).

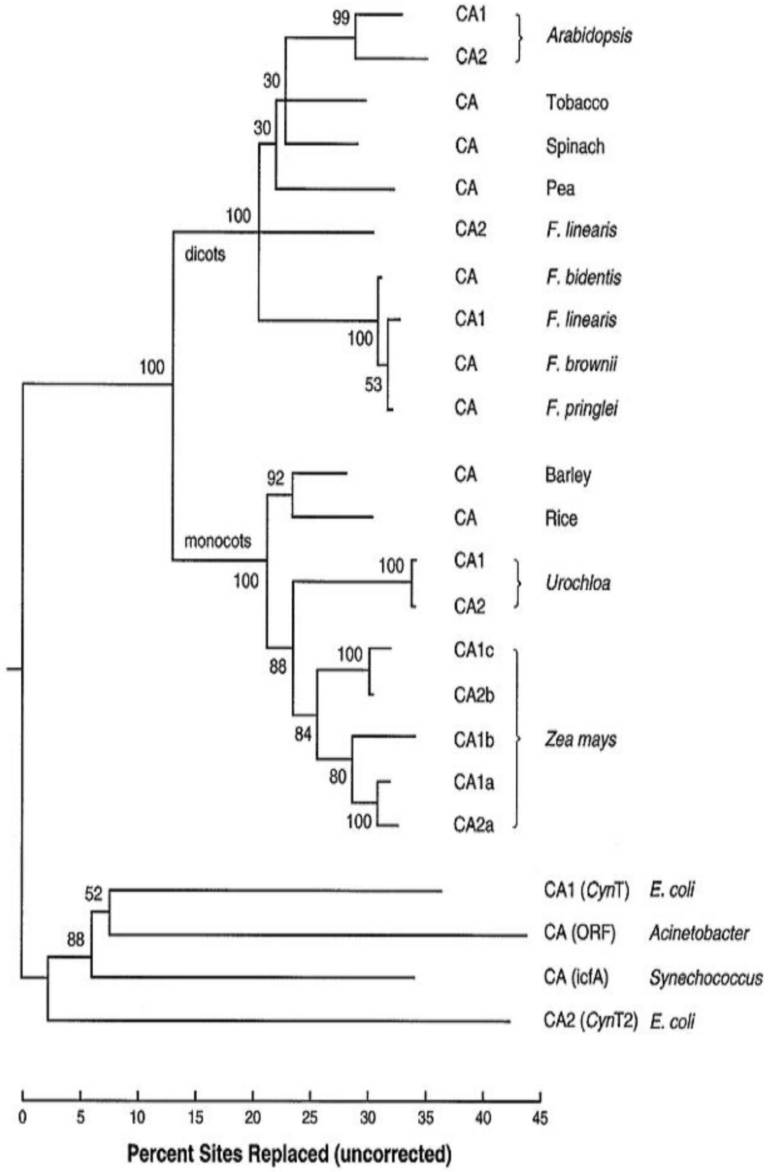


Fig. 1.3. *Phylogenetic analysis of the beta-CA gene family (Hewett-Emmett and Tashian, 1996).*

```
ArabCA1         MSTAPLSGFFLTSLSPSQSSLQKLSLRTSSTVACLPPASSSSSSSSSSSSRSVPTLIRNE 60
ArabCA2         ------------------------------------------------------------
Tobacco         MSTASINS-CLT-ISPAQASLKKPT--RPVAFARLS---------NSSSSTSVPSLIRNE 47
Spinach         MS--TING-CLTSISPSRTQLKNTSTLRPTFIANSR---------VNPSSSVPPSLIRNQ 48
Pea             MSTSSINGFSLSSLSPAKTSTKRTT-LRPFVSASLN-------TSSSSSSSTFPSLIQDK 52
MaizeCA3a       MYTLPVRATTSSIVP------------ACHPRAVLLL-----------------RLR--- 28
MaizeCA2a       MYTLPVRATTSSIVASLATPAPSSSSGSGRPRLRLIRNAPVFAAPATVCKRDGGQLRSQT 60
MaizeCA3b       ------------------------------------------------------------
MaizeCA3c       ------------------------------------------------------------
MaizeCA2c       ------------------------------------------------------------
Rice            MSTAAAAAAAQSWCFATVTPRSRATVVASLASP--------------------------- 33
E.coli          ------------------------------------------------------------


ArabCA1         P----------VFAAPAPIIAPYWSEEMGTEAYDEAIEALKKLLIEKEELKTVAAAKVEQ 110
ArabCA2         ---------------------------MGNESYEDAIEALKKLLIEKDDLKDVAAAKVKK 33
Tobacco         P----------VFAAPTPIINPILREEMAKESYEQAIAALEKLLSEKGELGPIAAARVDQ 97
Spinach         P----------VFAAPAPIITPTLKEDMA---YEEAIAALKKLLSEKGELENEAASKVAQ 95
Pea             P----------VFASSSPIITPVLREEMG-KGYDEAIEELQKLLREKTELKATAAEKVEQ 101
MaizeCA3a       -----------PPGSG-------------------------------SSG---------- 36
MaizeCA2a       REIERERKGGHPPAGGHKRGGERGQRRGGEEEEDEQLPLPSEKKGGASEGEAVHRYPHLV 120
MaizeCA3b       ------------------------------------------------------------
MaizeCA3c       ------------------------------------------------------------
MaizeCA2c       ------------------------------------------------------------
Rice            -----------SPSSS------------------------SSSSSNSSNLPAPFRPRLIR 58
E.coli          ------------------------------------------------------------


ArabCA1         ITAALQTGTSSD-------KKAFDPVETIKQGFIKFKKEKYETNPALYGELAKGQSPKYM 163
ArabCA2         ITAELQAASSSD-------SKSFDPVERIKEGFVTFKKEKYETNPALYGELAKGQSPKYM 86
Tobacco         ITAELQSSDGS---------KPFDPVEHMKAGFIHFKTEKYEKNPALYGELSKGQSPKFM 148
Spinach         ITSELADGGTP---------SASYPVQRIKEGFIKFKKEKYEKNPALYGELSKGQAPKFM 146
Pea             ITAQLGTTSSSDG------IPKSEASERIKTGFLHFKKEKYDKNPALYGELAKGQSPPFM 155
MaizeCA3a       TPR----LRRP-----ATVVGMDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYM 87
MaizeCA2a       TPSEPEALQPPPPPSKASSKGMDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYM 180
MaizeCA3b       ----------P-----------QDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYM 39
MaizeCA3c       ----------P-----------QDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYM 39
MaizeCA2c       ----------P-----------QDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYM 39
Rice            NTPVFAAPVAP--------AAMDAAVDRLKDGFAKFKTEFYDKKPELFEPLKAGQAPKYM 110
E.coli          -------------------------MKEIIDGFLKFQREAFPKREALFKQLATQQSPRTL 35


ArabCA1         VFACSDSRVCPSHVLDFQPGDAFVVRNIANMVPPFDKVKYGGVGAAIEYAVLHLKVENIV 223
ArabCA2         VFACSDSRVCPSHVLDFHPGDAFVVRNIANMVPPFDKVKYAGVGAAIEYAVLHLKVENIV 146
Tobacco         VFACSDSRVCPSHVLNFQPGEAFVVRNIANMVPAYDKTRYSGVGAAIEYAVLHLKVENIV 208
Spinach         VFACSDSRVCPSHVLDFQPGEAFMVRNIANMVPVFDKDKYAGVGAAIEYAVLHLKVENIV 206
Pea             VFACSDSRVCPSHVLDFQPGEAFVVRNVANLVPPYDQAKYAGTGAAIEYAVLHLKVSNIV 215
MaizeCA3a       VFACSDSRVCPSVTLGLQPGEAFTVRNIASMVPPYDKIKYAGTGSAIEYAVCALKVQVIV 147
MaizeCA2a       VFACSDSRVCPSVTLGLQPGEAFTVRNIASMVPPYDKIKYAGTGSAIEYAVCALKVQVIV 240
MaizeCA3b       VFACSDSRVCPSVTLGLQPAKAFTVRNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLV 99
MaizeCA3c       VFACSDSRVSPSVTLGLQPGEAFTVRNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLV 99
MaizeCA2c       VFACSDSRVCPSVTLGLQPGEAFTVRNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLV 99
Rice            VFSCADSRVCPSVTMGLEPGEAFTVRNIANMVPAYCKIKHAGVGSAIEYAVCALKVELIV 170
E.coli          FISCSDSRLVPELVTQREPGDLFVIRNAGNIVPSYGP-EPGGVSASVEYAVAALRVSDIV 94


ArabCA1         VIGHSACGGIKGLMSFPLDGNNSTDFIEDWVKICLPAKSKVISELGDSAFEDQCGRCERE 283
ArabCA2         VIGHSACGGIKGLMSFPLDGNNSTDFIEDWVKICLPAKSKVLAESESSAFEDQCGRCERE 206
Tobacco         VIGHSACGGIKGLMSLPADGSESTAFIEDWVKIGLPAKAKVQGEHVDKCFADQCTACEKE 268
Spinach         VIGHSACGGIKGLMSFPDAGPTTTDFIEDWVKICLPAKHKVLAEHGNATFAEQCTHCEKE 266
Pea             VIGHSACGGIKGLLSFPFDGTYSTDFIEEWVKIGLPAKAKVKAQHGDAPFAELCTHCEKE 275
MaizeCA3a       VIGHSCCGGIRALLSLKDGAPDNFTFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKE 207
MaizeCA2a       VIGHSCCGGIRALLSLKDGAPDNFTFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKE 300
MaizeCA3b       VIGHSCCGGIRALLSLKDGAPDNFHFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKE 159
MaizeCA3c       VIGHSCCGGIRALLSLQDGAPDTFHFVEDWVKIAFIAKMKVKKEHASVPFDDQWSILEKE 159
MaizeCA2c       VIGHSCCGGIRALLSLQDGAAYTFHFVEDWVKIGFIAKMKVKKEHASVPFDDQCSILEKE 159
Rice            VIGHSRCGGIKALLSLKDGAPDSFHFVEDWVRTGFPAKKKVQTEHASLPFDDQCAILEKE 230
E.coli          ICGHSNCGAMTAIASCQC--MDHMPAVSHWLRYADSAR-VVNEARPHSDLPSKAAAMVRE 151
```

```
ArabCA1         AVNVSLANLLTYPFVREGLVKGTLALKGGYYDFVKGAFELWGLEFGLSETSSVKDVATIL 343
ArabCA2         AVNVSLANLLTYPFVREGVVKGTLALKGGYYDFVNGSFELWELQFGISPVHSI------- 259
Tobacco         AVNVSLGNLLTYPFVREGLVKKTLALKGGHYDFVNGGFELWGLEFGLSPSLSV------- 321
Spinach         AVNVSLGNLLTYPFVRDGLVKKTLALQGGYYDFVNGSFELWGLEYGLSPSQSV------- 319
Pea             AVNASLGNLLTYPFVREGLVNKTLALKGGYYDFVKGSFELWGLEFGLSSTFSV------- 328
MaizeCA3a       AVNVSLQNLKSYPFVKEGLAGGTLKLVGAHYSFVKGQFVTWEP----------------- 250
MaizeCA2a       AVNVSLQNLKSYPFVKEGLAGGTLKLVGAHSHFVKGQFVTWEP----------------- 343
MaizeCA3b       AVNVSLQNLKSYPLVKEGLAGGTSSGW-PHYDFVKGQFVTWEP----------------- 201
MaizeCA3c       AVNVSLENLKTYPFVKEGLANGTLKLIGAHYDFVSGEFLTWKK----------------- 202
MaizeCA2c       AVNVSLENLKTYPFVKEGLANGTLKLIGAHYDFVSGEFLTWKK----------------- 202
Rice            AVNQSLENLKTYPFVKEGIANGTLKLVGGHYDFVSGNLDLWEP----------------- 273
E.coli          NVIAQLANLQTHPSVRLALEEGRIALHGWVYDIESGSIAAFDGATRQFVPLAANPRVCAI 211

ArabCA1         HWKL---- 347
ArabCA2         --------
Tobacco         --------
Spinach         --------
Pea             --------
MaizeCA3a       --------
MaizeCA2a       --------
MaizeCA3b       --------
MaizeCA3c       --------
MaizeCA2c       --------
Rice            --------
E.coli          PLRQPTAA 219
```

Fig 1.4. *Amino acid alignment of beta-CA sequences (Roeske and Ogren, 1990; Majeau and Coleman, 1992; Suzuki and Burnell, 1995; Burnell* et al.*, 1990a; Burnell and Ludwig, 1997).*

The differences between monocot and dicot beta-CAs include differences in molecular weight, quaternary structure and sensitivity to inhibitors. In addition, immunological experiments have indicated that there are antigenic differences between CAs from monocot and dicot species. Mono-specific antibodies raised against dicot CA from *Spinacia oleracea* (spinach) were more reactive against CA from dicot plant extracts than from monocot plant extracts (Okabe *et al.,* 1984). In contrast, antibodies raised against monocot CA from maize were cross-reactive with leaf extracts from a variety of other monocot species, but only quantitatively titrated CA activity from monocot plant extracts (Burnell, 1990). Monocot CAs were first reported as having a molecular weight of approximately 40 kDa and were predominantly monomers, which was similar to the animal CAs. The dicot CAs were approximately 180 kDa, with the association of several subunits required to form active enzyme (Atkins *et al.,* 1972a). The complexity of differences between the monocot and dicot CAs in higher plants is further compounded by the photosynthetic mechanism that these plants employ in order to produce carbon compounds.

Gamma CA gene family

An active gamma-CA was first identified in the archaeon *Methanosarcina thermophila,* and thought to be involved in acetate catabolism (Iverson *et al.,* 2000). The crystal structure was resolved indicating that the gamma-CA was a homo-trimer with a similar active-site structure to alpha-CAs. The active-site structure was also investigated using extended X-ray absorption fine structure (EXAFS) and while the catalytic zinc ion is coordinated by three histidine residues, these have different spacing in the linear sequence than for alpha-CAs (Hewett-Emmett and Tashian, 1996; Alber *et al.,* 1999; Moroney *et al.,* 2001). There are two gamma-CAs in *E. coli,* which also contains two beta-CA isoforms (Tripp *et al.,* 2004). All three CA gene families have been identified in Arabidopsis, including five gamma-CAs (Parisi *et al.,* 2004). The role of gamma-CAs in Arabidopsis was clarified by the generation of homozygous knockout mutant strains. A decrease in the abundance of Complex I of the mitochondrial respiratory chain in the knockout mutant strains indicated that all five of the gamma-CA isoforms contribute to complex assembly and are associated with Complex I (Perales *et al.,* 2005; Sunderhaus *et al.,* 2006). Additionally, the gamma-CAs were down-regulated when the plant was grown in elevated atmospheric carbon dioxide conditions, suggesting gamma-CAs remove excess carbon dioxide that is produced during photorespiration.

Delta CA gene family

The fourth CA gene family was identified in marine phytoplankton, though these species also contain CA representatives from all five gene families. The first delta-CA was found in the marine diatom, *T. weissflogii* (McGinn and Morel, 2008). The crystal structure of this enzyme has been resolved, and while containing no sequence homology with the alpha-CAs, has a similar active site structure with the active site zinc ion coordinated by the imidazole groups of three histidine residues. A delta-CA has also been identified in the marine coccolithophorid *Emiliania huxleyi,* which was distinguished from the gamma-CA also present in this species (Soto *et al.,* 2006). A third delta-CA has been characterized in the dinoflagellate *Lingulodinium polyedrum* and is associated with the plasma membrane improving carbon dioxide availability (Lapointe *et al.,* 2008). All three delta-CAs are involved in carbon metabolism and the CCM in these marine organisms, providing carbon dioxide for Rubisco.

Epsilon CA gene family

T. weissflogii contains a delta-CA as well as a cadmium-dependent epsilon-CA, which is the first example of a cadmium-containing metallo-enzyme (Lane *et al.,* 2005). Cadmium and zinc are both group II transition metals and have eight electrons, and it is likely the active site structure of CA could contain a cadmium molecule rather than zinc. However, cadmium is a heavy metal that is both toxic to the organism, in this case the freshwater macrophyte *Ceratophyllum demersum,* and detrimental to CA activity (Aravind and Prasad, 2004). CA activity is negatively affected by peroxidation and protein fragmentation resulting from the production of reactive oxygen species. In the presence of zinc, the sulphydryl groups of CA would be protected from oxidation. The crystal structure of the epsilon-CA from *Halothiobacillus neapolitanus* has recently been resolved, and analysis of this structure indicated that the enzyme was actually a variant beta-CA, suggesting the epsilon classification should be withdrawn (Fabre *et al.,* 2007).

Photosynthesis

Sunlight is the ultimate energy source for biological processes on earth and photosynthesis is the mechanism whereby plants, algae and some microorganisms convert light energy from the sun into chemical energy (Lawlor, 2001). Many physical and chemical reactions are involved resulting in the assimilation of carbon from carbon dioxide in the atmosphere, for the production of carbohydrates and the production of oxygen from water (Fig. 1.5).

$$6CO_2 + 6H_2O \xrightarrow{\textit{Light energy}} C_6H_{12}O_6 + 6O_2$$

Fig. 1.5. *The overall reaction of photosynthesis, showing the conversion of carbon dioxide to carbohydrates, using water and light energy from the sun, while producing oxygen.*

The light-dependent reactions of photosynthesis result in the conversion of light energy (400-700 nm) into chemical energy in the form of ATP and the reducing compound NADPH, which provide the energy for biosynthesis within a cell. The light-independent reactions of photosynthesis include the Calvin-Benson cycle, also called the photosynthetic carbon reduction cycle, which generates carbohydrates by fixing atmospheric carbon using energy and reducing power harvested in the light-dependent reactions (Mathews and Van Holde, 1996). Both the light-dependent and light-independent reactions occur in the chloroplast. Photosynthetic carbon assimilation depends on Rubisco, a large multi-subunit enzyme that is responsible for the initial fixation of carbon dioxide into a sugar molecule through ribulose-1,5-bisphosphate (RuBP) carboxylation (Lorimer, 1981).

Rubisco is a very inefficient catalyst lacking substrate specificity, and this limits the efficiency of carbon fixation. Rubisco uses both carbon dioxide and oxygen as substrates to function as either a carboxylase or an oxygenase (Lorimer, 1981). The reaction of RuBP with carbon dioxide is productive and leads to photosynthetic carbon reduction. In contrast, the reaction of RuBP with oxygen leads to the release of carbon dioxide through photorespiration, which is energetically wasteful (Edwards *et al.,* 2003). As a result, this system is only effective when the concentration of carbon is high. Therefore high levels of atmospheric oxygen would have been favourable for the evolution of a CCM, and in particular the C_4 pathway, which is an organic carbon pump (Fig. 1.6). It has been reported that in response to strong selective pressures the C_4 pathway of photosynthesis independently evolved over 45 times in 19 families of angiosperms, and thus represents one of the most convergent of evolutionary phenomena (Sage, 2004).

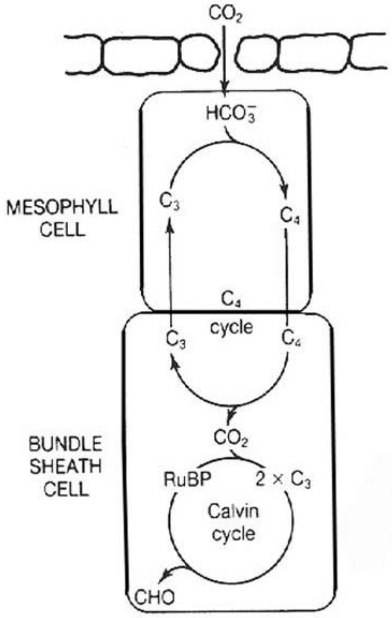


Fig. 1.6. *Depiction of the C_4 photosynthetic pathway, showing anatomical compartmentalization of carbon assimilation and fixation that occurs in some C_4 plant species (Lea and Leegood, 1999).*

1.2.1 C_4 Photosynthesis

The C_4 photosynthetic pathway was discovered in experiments that showed that in some plant species radioactively labelled carbon from carbon dioxide was initially incorporated into four-carbon compounds such as oxaloacetate (Hatch *et al.,* 1967). The C_4

pathway is usually a dual-celled system allowing for the separation of the processes that occur in the external mesophyll cells and the inner bundle sheath cells that surround the vascular tissues (Edwards *et al.,* 1985). This system is referred to as Kranz anatomy; although there is considerable variation in Kranz anatomy amongst C_4 plants (Edwards *et al.,* 2003). The general pattern is that of an outer layer derived from mesophyll cells, in which phospho*enol*pyruvate (PEP) carboxylase is located and thus initial carboxylation occurs (Wu and Wedding, 1994). The inner layer consists of the bundle sheath cells in which the enzymes of the Calvin-Benson cycle are located, and can be derived from any number of cell layers that are near or within the vascular bundle (Sage, 2004). The different cell types associated with the C_4 photosynthetic pathway allows for separation of cellular activities (Fig. 1.6). After the CA-catalyzed conversion of carbon dioxide to bicarbonate in the mesophyll cells, the primary carbon fixation reaction occurs, catalyzed by PEP carboxylase, producing oxaloacetate, a four-carbon compound. Oxaloacetate can be converted to other four carbon compounds (aspartate and malate) which then diffuse to the bundle sheath cells via plasmodesmata (Weiner *et al.,* 1988).

The efficiency of the C_4 pathway is an important factor for agronomically important species such as maize and *Saccharum officinarum* (sugarcane), where plant yields are increased. As a consequence of this the genetic transfer of C_4 traits to C_3 plants has become a desirable outcome for improving crop yields in C_3 species (see Sheehy *et al.,* 2007). C_4 plants achieve increased photosynthetic production due to the concentration of carbon dioxide around Rubisco, which limits the oxygenation reaction that leads to photorespiration, the production of phosphoglycolate and carbon dioxide loss. Thus an increase in photosynthesis efficiency in C_3 plants could be achieved by either reducing the degree of energetic loss resulting from photorespiration, or by introducing C_4 enzymes to increase the concentration of carbon dioxide near Rubisco.

The effects of photorespiration were reduced when five enzymes of the *E. coli* glycolate catabolism pathway were transferred into the C_3 plant Arabidopsis (Kebeish *et al.,* 2007). This pathway uses glycolate that is produced by the Rubisco-catalyzed oxygenation reaction to produce carbon dioxide in the vicinity of Rubisco, catalyzed by glyoxylate carboligase. This pathway also regenerates the carboxylation substrate 3-phosphoglycerate, with the end result being increased plant growth and biomass. Alternatively, increased expression of a C_4 enzyme does appear to alter carbon metabolism in C_3 plants. For example, Rubisco over-expression leads to increased photosynthetic capability in rice, but only at low temperatures and with a resulting decrease in nitrogen availability (Makino and Sage, 2007). Whether the separation of the processes of carbon assimilation and fixation is required to increase the concentration of carbon dioxide near Rubisco, thereby improving Rubisco efficiency, remains to be determined. This is achieved by Kranz anatomy in C_4 plants, although there are several examples of single-celled C_4 mechanisms (Khan, 2007). An additional alternative is the over-expression of multiple C_4 enzymes in C_3 plants in order to improve photosynthetic yields (Miyao, 2003).

CA in C_3 plants

Most biochemical studies have been conducted on CA from C_3 dicot plants, where the majority of CA has been located in the chloroplast stroma of mesophyll cells, accounting for up to 2% of total leaf protein (Okabe *et al.,* 1984; Björkbacka *et al.,* 1999). Rubisco is the primary carbon-fixing enzyme in the chloroplasts of C_3 plants and the supply of carbon dioxide to Rubisco is facilitated by CA (Sultemeyer *et al.,* 1993). CA expression co-localizes with Rubisco expression due to the inorganic carbon requirements of the cell (Majeau *et al.*, 1994). CA is encoded in the nucleus and transport into the chloroplast is dependent upon the presence of a transit peptide. At least two isozymes of CA have been detected in a range of C_3 plants (Atkins *et al.,* 1972a). The second CA isozyme locates to the cytosol where it aids diffusion of carbon dioxide across the cell membrane (Reed and Graham, 1980).

One of the earliest CA enzymes discovered in plants was purified from the leaves of pea*,* a C_3 dicot. Initially it was thought to associate as a hexamer containing six zinc atoms, though it was later shown to be an octamer through structural X-ray analysis (Kisiel and Graf, 1972; Moroney *et al.,* 2001). In another C_3 dicot, spinach, CA is a hexamer containing six zinc atoms with a subunit size of approximately 35 kDa (Pocker and Ng, 1973; Fawcett *et al.,* 1990). In these plants CA was suggested as being essential for the maximum rate of photosynthesis (Ohki, 1976).

In the dicot *P. vulgaris* there was a correlation between CA activity and zinc deficiency, while other enzymes including Rubisco, glycolic oxidase and malic enzyme were not affected (Edwards and Mohamed, 1973). A zinc ion is required for CA catalysis, and CA activity had been used to indicate zinc deficiency (Bar-Akiva *et al.,* 1971). Despite decreased CA activity, photosynthetic rates were not changed as evidenced by models of inorganic carbon flow (Graham *et al.,* 1984). Nevertheless, stomatal conductance was increased in *Nicotiana tabacum* (tobacco) plants transformed with an anti-sense CA construct, which decreased chloroplastic CA to 1% of wild type levels (Majeau *et al.,* 1994). Stomatal conductance increased to raise the intercellular carbon dioxide concentration, compensating for the loss of CA activity.

The active site structure of spinach and pea CA was analyzed by EXAFS and kinetic studies. A cysteine-histidine-cysteine-water ligand scheme was proposed for coordination of the zinc ion at the active site (Rowlett *et al.,* 1994). This was confirmed by site-directed mutagenesis of spinach CA (Bracey *et al.,* 1994). However, X-ray structures of the CAs from *P. purpureum* and *E. coli* indicated that instead of a water molecule, the fourth zinc-binding ligand is an asparagine residue (Rowlett *et al.,* 2002). The active-site histidine may not be required for catalysis, although it is most likely in the vicinity of the active site. CA mutation studies in Arabidopsis showed that the histidine residue is essential for a proton-shuttling function. Substituting histidine with alanine at this position altered the kinetic values for the enzyme, and only imidazole-type buffers were able to replace the proton-shuttling function of histidine (Björkbacka *et al.,* 1999).

CA is also involved in non-photosynthetic biochemical pathways in C_3 plants. In tobacco, a chloroplast CA was identified as the salicylic acid-binding protein 3 (SABP3), implying a role in the hypersensitive defense response (Slaymaker *et al.,* 2002). CA may also play a role in post-germinative growth of *Gossypium hirsutum* (cotton) seedlings, indicated by transcript abundance in this part of the plant (Hoang *et al.,* 1999). In *Medicago sativa* (alfalfa) symbiotic nitrogen-fixing nodules, two sources of CA activity have been reported. It is hypothesized that the first CA provides inorganic carbon substrates for PEP carboxylase, enabling production of carbon skeletons for ammonia assimilation, and the second is involved in gas exchange (Gàlvez *et al.,* 2000).

CA in C_4 plants

Plants that use the C_4 pathway of photosynthesis use a CCM to increase the concentration of carbon dioxide in the vicinity of Rubisco, made possible by Kranz anatomy. PEP carboxylase uses bicarbonate as a substrate to catalyze the first carbon-fixing reaction, which occurs in the mesophyll cells (Sheen and Bogorad, 1987; Jenkins *et al.,* 1987). Therefore for this reaction to occur, carbon dioxide that has diffused through the cell membrane must be converted to bicarbonate. This is catalyzed by CA located in the cytosol of mesophyll cells of C_4 plants, and is the first step of the C_4 photosynthetic pathway (Hatch and Burnell, 1990).

Initial reports indicated that there was less CA activity in C_4 plants than in C_3 plants, and that this activity was located in the cytosol of C_4 plants, but in the chloroplasts of C_3 plants (Everson and Slack, 1968; Poincelot, 1972; Triolo *et al.,* 1974). Analysis of the differential accumulation of proteins to the mesophyll or bundle sheath cells in the C_4 monocot, *Sorghum bicolor* (sorghum), confirmed expression of CA predominantly in the mesophyll cells (Wyrich *et al.,* 1998). CA has been purified from the mesophyll cell cytosol of the C_4 species *Amaranthus cruentus.* A second CA isozyme was proposed to be associated with the chloroplast membrane of the bundle sheath cells, where low levels of CA activity were detected (Guliev *et al.,* 2003). Initially it was suggested that CA activity in this part of the cell would inhibit efficient functioning of the C_4 pathway, and that loss of CA activity from bundle sheath chloroplasts was a necessary step in the evolution of C_4 photosynthesis (Burnell and Hatch, 1988). CA activity in the chloroplast would result in the conversion of carbon dioxide to bicarbonate decreasing the available carbon dioxide for Rubisco. In the light, the chloroplast stroma is approximately pH 8, which would result in the bicarbonate concentration being 50 times the carbon dioxide concentration at equilibrium (Burnell and Hatch, 1988).

The effects of CA activity in the bundle sheath cell cytosol were determined when *F. bidentis* was transformed with tobacco CA under the control of constitutive promoters. These plants displayed increased levels of photorespiration, and results consistent with bicarbonate leakage from the bundle sheath cells, indicating that CA expression in this cell type was detrimental (Ludwig *et al.,* 1998). Despite this, CA cDNA from *F. bidentis* has a leader sequence analogous to a chloroplast transit peptide (Cavallero *et al.,* 1994). Recently, three beta-CA isoforms have been identified in *F. bidentis,* which are encoded by a small multi-gene family (Tetu *et al.,* 2007). One of these was imported into chloroplasts in *in vivo* uptake assays, and was detected in chloroplasts by immuno-localization. This isozyme may be involved in non-photosynthetic biochemical processes such as lipid biosynthesis and antioxidant activity, similar to CA in C_3 plants.

The first evidence for CA association with the plasma membrane was in maize, where treatment with a detergent increased the CA activity recovered in crude plant extracts (Atkins *et al.,* 1972a). Using plasma membranes isolated by aqueous two-phase partitioning, approximately 35% of total cellular CA activity was found on the cytoplasmic side of the membrane (Utsunomiya and Muto, 1993). However the differences in the sub-cellular localisation of CA in C_3 and C_4 plants was highlighted by the finding that less CA activity was in the plasma membranes of the C_3 species examined.

Two hypotheses for the function of CA in the plasma membrane of C_4 plants have been proposed. The first predicts that the activity of CA in the mesophyll cells is equally dispersed around the cell, where it can provide bicarbonate for PEP carboxylase as well as facilitating the diffusion of carbon dioxide into the cell. The second hypothesis proposes that CA converts carbon dioxide leaking from the bundle sheath cells into bicarbonate, hence trapping bicarbonate in the mesophyll cells where it can be used by PEP carboxylase. For this to occur, CA in the plasma membrane must be located in close vicinity to the plasmodesmata that join the mesophyll and bundle sheath cells and this is yet to be experimentally shown (Burnell, 2000).

CA in NADP-ME type monocot C_4 plants

Maize as well as sorghum and sugarcane decarboxylate four carbon acids via NADP-malic enzyme (NADP-ME) located in the chloroplasts of bundle sheath cells and are thus classified as C_4 NADP-ME monocot plants. In this specific plant subtype beta-CA isozymes exist that differ significantly from all other reported forms of the enzyme due to transcript length and the presence of repeating sequences within the cDNA sequence (Burnell and Ludwig, 1997).

Hybridization of RNA from maize and sorghum with a beta-CA cDNA probe identified three transcripts that were approximately 1.5 kb, 1.9 kb, and 2.2 kb (Burnell and Ludwig, 1997; Wyrich *et al.,* 1998). In maize, the corresponding cDNA sequences have been designated CA1, CA2 and CA3, respectively (Burnell and Ludwig, 1997; Burnell, unpublished). All three CA cDNAs are composed of repeated sequences that are approximately 600 bp in length, and this primary structure is unique to NADP-ME C_4 monocots. CA1 and CA2 are composed of two repeat sequences, and CA2 has an additional 5′-leader sequence, including a unique 276 bp insert. CA3 has three repeat sequences (Fig. 1.7). Other examples of CAs that are composed of repeated sequences include the unicellular red alga *P. purpureum* that appears to be the product of gene duplication, and the alpha-CAs from *D. salina* and *T. gigas,* which have two active sites (Mitsuhashi *et al.,* 2000; Moroney *et al.,* 2001; Leggat *et al.,* 2005).

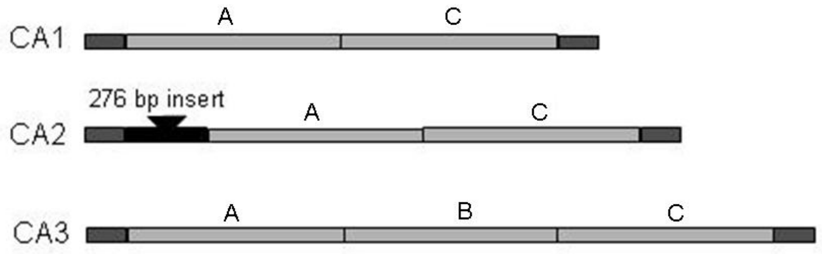


Fig. 1.7. *Schematic representation of the three beta-CA cDNAs from maize (Burnell and Ludwig, 1997; Burnell, unpublished). The repeat sequences are approximately 600 bp and are labelled A, B and C. The black shaded box represents the 276 bp insert unique to CA2. The dark grey shaded boxes represent the 5′-leader sequences and 3′-untranslated regions.*

The repeat sequences contain significant homology to each other, and are homologous to other monocot CA cDNA sequences (Fig. 1.4). Repeat B of CA3 may be a chimeric gene product, as the 5′-end of Repeat B has homology with Repeat C, and the 3′-end of Repeat B has homology with Repeat A (Hewett-Emmett and Tashian, 1996). There are no obvious processing sites in the nucleotide or amino acid sequences at the junctions of the repeats.

Several mechanisms for regulation of CA expression levels in maize leaf tissue have been examined. There is a positive correlation between mRNA levels of CA (and PEP carboxylase) and nitrogen availability, dependent on the presence of cytokinins (Sugiharto *et al.,* 1992a). Cytokinins are made in the root tissue of plants and stimulate transcription of CA in response to favourable environmental conditions for plant growth, such as when nitrogen is available. When the production of glutamine, a positive signal for nitrogen availability, was prevented by glutamine synthetase inhibition in detached maize leaves CA and PEP carboxylase transcripts decreased (Sugiharto *et al.,* 1992b). Light also positively regulates CA transcription, and CA mRNA levels increase when the leaf is illuminated (Burnell and Ludwig, 1997).

Immunological studies using a crude maize leaf protein extract have shown four immuno-reactive protein bands with molecular weights of 27, 28, 47 and 52 kDa, with the 47 kDa species being the most abundant (Burnell and Ludwig, 1997). The intensity of the 52 kDa band decreased in the absence of protease inhibitors, while the 28 kDa band increased in the presence of a detergent. Detergents such as Triton X-100 enable extraction of membrane-associated proteins and a CA isoform was recovered using this treatment (Utsunomiya and Muto, 1993). The molecular weight of the 27 kDa and 47 kDa species was 180 kDa determined by column chromatography (Burnell and Ludwig, 1997). This implies that the 27 kDa CA isoform associated at least as a hexamer, while the 47 kDa isoform was most likely a tetramer. This is consistent with other reported tertiary structures for beta-CAs, which include octomers, hexamers, tetramers and dimers (Hiltonen *et al.,* 1998; Kimber and Pai, 2000).

In sorghum there are three CA transcripts that are encoded by multiple genes. The 3′-untranslated regions of cDNA sequences obtained from library screening revealed differences in nucleotide sequence that enabled the generation of gene-specific probes. These were used in a Southern analysis of genomic DNA that demonstrated that the three CA isozymes were encoded by two different genes (Wyrich *et al.,* 1998). Approximately 7.5 kb of a beta-CA gene from maize has been sequenced and contains many small exons characteristic of plant genes (Burnell, unpublished). There is a long intron between Exon 1 and Exon 2, which may be involved in regulating the differential expression that results in the different CA isozymes observed in plant extracts. The longest exon in the gene is Exon 2 which corresponds to the 276 bp insert of CA2 (Burnell, 2000). Whether there is more than one gene encoding CA in maize is unknown, although the identification of a small multi-gene family in *F. bidentis* and the discovery of at least two CA genes in sorghum, also a C_4 NADP-ME type monocot, is highly suggestive that maize CAs are encoded by a multi-gene family (Tetu *et al.,* 2007; Wyrich *et al.,* 1998).

Summary

Maize, as well as sorghum and sugarcane, decarboxylate four carbon acids via NADP-ME located in the chloroplast of bundle sheath cells, and are thus classified as C_4 NADP-ME monocot plants. The CA isozymes from this particular group of C_4 plants are encoded by comparatively large transcripts, due to the presence of repeating sequences (Burnell and Ludwig, 1997; Wyrich *et al.,* 1998). These isozymes remain relatively uncharacterized. Immunological studies using a maize leaf extract have shown four protein bands with molecular weights of 27, 28, 47 and 52 kDa, none of which coincide with the size of proteins predicted by translation of the open reading frames of the cDNA sequences (Burnell and Ludwig, 1997). Therefore, the number of genes encoding beta-CA in maize, the mechanism by which the isozymes are expressed, as well as the structure and function of CA in maize, was the focus of this investigation.

CHAPTER 2

2. Materials and Methods

Molecular biology methods

2.1.1 Restriction endonuclease treatment

Reactions included up to 1 µg of DNA per 10 units of restriction endonuclease (Promega, Pharmacia Biotech, New England BioLabs) in the recommended buffer in a final volume of 20 µl. BSA (Promega, Sigma) was added where required to a final concentration of 100 µg.ml^{-1}. Reactions were incubated at the recommended temperature for the enzyme, usually 37°C, for 2-3 h in a heat block. Reactions were stopped through either heat inactivation at 65°C for 20 min or by the addition of 5 µl of Stop Mix (60 mM EDTA, 10 mM Tris-HCl pH 8, 30% v/v glycerol, 0.01% w/v bromophenol blue).

2.1.2 Agarose gel electrophoresis

Agarose (Sigma, Amresco) was dissolved in TBE buffer (90 mM Tris-HCl pH 8.3, 3 mM EDTA, 90 mM boric acid) at concentrations between 0.8-2% (w/v), and 5 µl of ethidium bromide (10 mg.ml^{-1}) added before the gel was cast. Electrophoresis was performed at 100 V for 20-40 min and the DNA visualised using a UV trans-illuminator. The size of linear double stranded DNA fragments was estimated from comparison with the migration of a molecular weight marker (Fermentas, New England BioLabs, Promega).

2.1.3 Purification of DNA from agarose gels

Two methods were used to purify DNA from an agarose gel, after a scalpel was used to excise the band. Purification of DNA was performed using the QIAGEN gel extraction kit and protocol, or alternatively the DNA fragment was isolated from the gel by dialysis. Cut dialysis tubing (Sigma) was prepared by boiling in 2% (w/v) sodium bicarbonate and 1 mM EDTA, pH 8 for 30 min. The tubing was washed with double distilled water and stored in 1 mM EDTA at 4°C. The agarose slice was sealed in the dialysis tubing with 300 µl of TBE buffer and submerged in the electrophoresis tank. Electro-elution was performed at 100 V for 20 min with a final reversal of the current for 10 s to release DNA from the tubing. The DNA was purified and concentrated by phenol/chloroform extraction and ethanol precipitation (Section 2.1.6). The DNA fragment was resuspended in 10-20 µl of preheated double distilled water/TE buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA).

2.1.4 Purification of total RNA from maize tissue

Approximately 100 mg of maize leaf/root tissue was ground with a mortar and pestle in liquid nitrogen. Total RNA was isolated from the ground tissue using the QIAGEN RNeasy Plant Mini Kit and protocol. The concentration of total RNA was measured spectrophotometrically (Section 2.1.5) and the quality of the RNA assessed by agarose gel electrophoresis (Section 2.1.2).

2.1.5 Determining the concentration of single and double stranded DNA, and RNA

DNA/RNA was quantified by spectrophotometric analysis using a Beckman DU 650 spectrophotometer. A wavelength scan from 200-300 nm was performed and the concentration calculated from the absorbance at 260 nm. The DNA/RNA quality was assessed by the A_{260}/A_{280} ratio. The concentration of DNA was calculated as 50 $\mu g.ml^{-1}$ when the A_{260} was equal to 1, and RNA or single stranded DNA as 40 $\mu g.ml^{-1}$ when the absorbance at 260 nm was 1 (Sambrook *et al.,* 1989).

2.1.6 Phenol/chloroform extraction and ethanol precipitation of DNA

Solutions were made up to 50 µl with double distilled water in a 1.5 ml microcentrifuge tube before the addition of an equal volume of phenol/chloroform (1:1). The mixture was vortexed and centrifuged for 10 min at 16,100 × *g*. The supernatant was transferred to a fresh 1.5 ml microcentrifuge tube and the DNA precipitated by the addition of 0.1x vol of 3 M sodium acetate, pH 5.2 and 2x vol cold 95% ethanol, and incubation at -20°C for at least 1 h. DNA was pelleted by centrifugation at 4°C for 20 min at 16,100 × *g*. The supernatant was removed and the pellet washed twice with 500 µl of cold 70% ethanol before being vacuum dried. The pellet was resuspended in 5-10 µl of preheated (65°C) distilled water/TE.

2.1.7 Preparation of competent cells

Chemically competent cells (*E. coli,* strain: NM522) were prepared using an adaptation of the method of Cohen *et al.*, 1972 (Sambrook *et al.*, 1989). A single colony was picked from a fresh LB agar plate that had been previously spread using a -80ºC glycerol stock, and used to inoculate 5 ml of LB media, shaken overnight at 37ºC. These cultures were used to inoculate 500 ml of LB media, shaken at 37ºC for 3-4 h, or until the OD_{600} was between 0.5-0.6. The cultures were cooled to 0ºC by incubation on ice for 30 min, and then harvested in pre-cooled 50 ml Falcon tubes, by centrifugation at 3,000 × *g* for 20 min at 2ºC. The pellets were resuspended in 25 ml ice cold 0.1 M $CaCl_2$ and incubated on ice for 30 min. Cells were recovered by centrifugation at 3,000 × *g* for 20 min at 2ºC and the cell pellet resuspended in 2 ml of ice cold 0.1 M $CaCl_2$ for each 50 ml of original culture. The suspension was then aliquoted (200 µl) into pre-chilled sterile 1.5 ml microcentrifuge tubes and snap-frozen in liquid nitrogen. Competent cells were stored at -80ºC until required.

2.1.8 Preparation of glycerol stocks

Glycerol stocks were prepared from 5 ml cultures that had been inoculated with a single colony and grown at 37°C overnight with shaking. To 500 µl of culture, 500 µl of sterile 80% (v/v) glycerol was added, mixed and the glycerol stock stored at -80ºC.

2.1.9 Transformation of competent cells and blue/white selection

An aliquot of competent cells were thawed from -80°C on ice. Plasmid DNA (either 10 ng purified plasmid DNA, or a 15 µl of ligation reaction, Section 2.1.11) was added and the cells incubated on ice for 20 min. Heat-shock was performed at 42°C for 90 s in a water bath, before the cells were returned to ice for 5 min. The transformation mixture was plated onto pre-warmed LB plates containing the appropriate antibiotic (Table 2.1), and incubated overnight at 37°C.

Table 2.1. Antibiotic concentrations used for culturing *E. coli* (Sambrook *et al.,* 1989).

	Stock Solution	**Working Solution**
Carbenicillin/Ampicillin	25 $mg.ml^{-1}$ in ddH_2O	50 $\mu g.ml^{-1}$
Chloramphenicol	25 $mg.ml^{-1}$ in ethanol	25 $\mu g.ml^{-1}$
Kanamycin	10 $mg.ml^{-1}$ in ddH_2O	10 $\mu g.ml^{-1}$
Tetracyclin	5 $mg.ml^{-1}$ in ethanol	10 $\mu g.ml^{-1}$

Blue/white selection was possible where the pGEM®-T vector system (Promega), or pBluescript (Stratagene) were used, as these vectors contain the *lacZ* coding region adjacent to the multiple cloning site (Appendix 4.1). LB plates were spread with 50 µl of 20 mM IPTG/2% X-Gal (in N, N,-dimethylformamide) solution and allowed to dry before plating of transformed cells.

2.1.10 Plasmid DNA preparation

Plasmid DNA was prepared from 5 ml LB cultures that had been inoculated with a single colony and shaken at 37ºC overnight. The cells were harvested by centrifugation at 5,292 × *g* for 10 min.

Purification was performed using the QIAGEN mini-prep kit and protocol. The plasmid DNA was eluted using 30 µl of either preheated (65ºC) distilled water or preheated elution buffer (QIAGEN) and by centrifugation at 16,100 × *g* for 60 s.

Alternatively, plasmid DNA was purified by the alkaline lysis method. After harvesting, cell pellets were resuspended in 100 µl GTE buffer (50 mM glucose, 25 mM Tris-HCl pH 8, 10 mM EDTA), and lysed with 100 µl lysis buffer (200 mM NaOH, 1% w/v SDS). Neutralization was achieved by the addition of 150 µl of 3 M potassium acetate, 2 M acetic acid pH 5.4 before centrifugation at 16,100 × *g* for 10 min. The plasmid DNA was precipitated by the addition of 1 ml of 95% ethanol to the supernatant and pelleted by centrifugation at 16,100 × *g* for 10 min. The DNA pellet was washed twice with cold 70% ethanol and resuspended in 30 µl of distilled water.

2.1.11 Ligation reactions

Ligation of plasmid vector with insert DNA was performed in 15 µl reactions, with a ratio of insert to vector of 3:1. The insert was either a PCR product or a restriction enzyme DNA fragment that had been excised from an agarose gel and purified using the QIAGEN gel extraction kit and protocol (Section 2.1.3). Reactions contained 1 unit of T4 DNA ligase and 1x vol T4 DNA ligase buffer (New England Biolabs, Promega).

Control reactions were also performed to gauge the effectiveness of alkaline phosphatase treatment (Section 2.1.12) of the vector where required, and to eliminate contamination. Reactions were incubated at 4°C for 24 h, 16°C overnight or at room temperature for 4 h. When incubated at room temperature the vector and insert DNA were first heated to 50°C for 1 min and then incubated at 37°C for 2 min before the enzyme and its buffer were added.

2.1.12 Alkaline phosphatase treatment of DNA fragments

Where necessary, the vector for a ligation reaction was treated with alkaline phosphatase. Reactions included 1 unit of shrimp alkaline phosphatase (SAP, Promega) per 0.5 pmol of DNA ends with 1x vol SAP buffer (Promega) in a reaction volume of 30 µl. Reactions were incubated at 37°C for 15 min, before the enzyme was heat inactivated at 95ºC for 5 min.

2.1.13 DNA Sequencing

Samples of approximately 1 µg of dried plasmid DNA were submitted to Macrogen Inc, Korea. Where necessary, 100 pmol of the appropriate primer were also dried and sent with the plasmid template. Sequence data was analysed using Sequencher and MacVector^{TM} 7.0 sequence analysis software.

2.1.14 Polymerase Chain Reaction (PCR)

Reactions were performed in a total volume of 20-50 µl (Table 2.2). Higher concentrations of primers and $MgCl_2$ were used when amplifying fragments from cDNA, genomic DNA or cDNA/genomic DNA libraries. The temperature cycles used depended on the melting temperature of the primers and the length of the product, but included denaturation (94/95ºC), annealing (45-65ºC) and extension steps (68-72ºC). Reactions were cycled on either a Thermo Hybaid PCR Express machine, or a MJ Research PTC-200 Peltier Thermal Cycler.

Table 2.2. Components of a standard PCR

Component	**Concentration**
Reaction buffer (Promega, Stratagene, Invitrogen)	1x
dNTPs (Promega)	0.2 mM
$MgCl_2$ (Promega)	1.5-2.5 mM
Primers	0.25-0.5 µM
Taq polymerase (Promega-*GoTaq*, Stratagene-*Pfu Turbo*, Invitrogen-*Platinum Pfx*)	1-1.5 units

Protein methods

2.2.1 Denaturing Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis (SDS-PAGE)

Proteins were separated by SDS-PAGE using BioRad vertical protein electrophoresis equipment. Gel percentage (10-15%) was calculated based on 5 ml gel volume and 40% (v/v) acrylamide (Bis/Acryl™ 29:1, 40% w/v, Amresco). Samples were mixed with an equal volume of cracking buffer (0.01% w/v bromophenol blue, 125 mM Tris-HCl pH 6.7, 2% w/v SDS, 10% v/v glycerol, 10% v/v 2-mercaptoethanol), and boiled for 5-10 min before loading. Electrophoresis was performed in running buffer (192 mM glycine, 25 mM Tris, 0.1% w/v SDS) at 100 V until the samples had reached the separating gel, then 200 V until the bromophenol blue dye front had reached the bottom of the gel.

Once run, the gel was stained in Coomassie blue staining solution (1.25% w/v Coomassie R-250/G-250, 50% v/v methanol, 10% v/v glacial acetic acid) for at least 1 h and then destained (20% v/v methanol, 20% v/v glacial acetic acid). Gels were dried in gel drying film (Promega) after washing in 20% (v/v) methanol for 20 min.

Analysis of protein subunit size involved comparison to the migration of a pre-stained molecular weight marker (Fermantas, Invitrogen, New England BioLabs).

2.2.2 Western blotting

After gel electrophoresis, the proteins were transferred from the gel to polyvinylidene fluoride (PVDF) membrane (Immobilon-P, 0.45 µm or Immobilon-P^{SQ} 0.2 µm, Millipore) using the BioRad Mini Trans-Blot system. The transfer was conducted at 100-200 V for 20-50 min between four layers of blotting paper and four layers of scourer pads, which had been equilibrated with transfer buffer (192 mM glycine, 25mM Tris, 20% v/v methanol, 0.1% w/v SDS). Before use, the PVDF membrane was washed with 100% methanol for 15 s.

To ensure complete transfer of protein from the gel and the membrane, the gel was stained with Coomassie blue post-transfer.

2.2.3 Ponceau staining

After transfer, the membrane was rinsed briefly with destaining solution and then submerged in Ponceau staining solution (3% w/v trichloroacetic acid, 3% w/v sulphosalicylic acid, 0.2% w/v Ponceau S) for 5-10 min. The membrane was then washed with destaining solution until protein bands became visible. The membrane was washed thoroughly with distilled water before immunoblotting.

2.2.4 Immunoblotting

The membrane was incubated with an appropriate blocking agent, depending on the primary antibody. Blocking occurred for 30-60 min in either 5% (w/v) milk powder or 3% (w/v) BSA (Sigma) in TBST (10mM Tris-HCl pH 7.5, 150 mM NaCl, 0.05% v/v Tween-20) with agitation. After blocking, the primary antibody was added at the appropriate dilution, sealed with the membrane in a plastic bag and incubated at room temperature for at least 3 h. After this time, the membrane was subjected to three 15 min washes in TBST. The secondary antibody was added at the appropriate dilution with TBST, sealed in a plastic bag with the membrane and incubated for at least 1 h. The membrane was again washed three times in TBST for 15 min. For colour development, a diaminobenzidine (DAB) tablet and a urea/hydrogen peroxide tablet (Sigma) were dissolved in 5 ml of distilled water before being poured over the membrane. Alternatively, 6 mg diaminobenzidine was dissolved in 9 ml of 10 mM Tris-HCl pH 7.5. 1 ml of 0.3% (w/v) $CoCl_2$ and 10 µl of hydrogen peroxide were added to this solution, which was then poured over the membrane for the colour development reaction. Colour development was terminated by washing the membrane with tap water.

2.2.5 Bradford assay for protein concentration

Protein concentrations were determined using the method of Bradford (Bradford 1976). Standard curves were prepared using 2-10 $\mu g.ml^{-1}$ of BSA (Promega) with 0.2x vol of assay dye reagent concentrate (Bio-Rad) and the solutions incubated for 5 min before absorbance readings were taken in triplicate at 595 nm.

Plant material

Maize seeds (*Zea mays,* Strain A188) were stimulated to germinate through submersion in water overnight. Approximately 25 seeds were planted per 25 cm diameter pot and grown in natural light and air conditions at temperatures between 16 to 27ºC for 4-8 weeks. Where necessary, leaf tissue was harvested, de-ribbed, vacuum-sealed and stored at -80ºC until required.

CHAPTER 3

3. Genomic Organization of the beta-CA Gene Family

Introduction

3.1.1 Maize genome complexity

The maize genome is approximately the same size as the human genome with a haploid content of 2.5 billion nucleotides on ten chromosomes. Only 7.5% of the maize genome is predicted to be gene-coding sequence in comparison with 25% of the human genome (Messing *et al.,* 2004; Venter *et al.,* 2001). The genome of the dicot model plant Arabidopsis was the first to be sequenced and has only 129 million nucleotides, which is significantly smaller than maize (Arabidopsis Genome Initiative, 2000). Rice, like maize, is a member of the grass family Poaceae and therefore has physiological similarity to maize, but has a genome size of 389 million nucleotides (International Rice Genome Sequencing Project, 2005). The genome size of sorghum is also significantly less than maize. With approximately 750 million nucleotides, the sorghum genome is only one third the size of the maize genome. As the divergence of maize and sorghum from a common ancestor occurred within the last 16 million years, the maize genome size must have increased after this time (Gaut and Doebley, 1997; Swigoňová *et al.,* 2004).

The increase in the size of the maize genome since the divergence of maize and sorghum has been attributed to the action of retro-transposons. It is predicted that up to 80% of the maize genome is made up of either high or low copy-number retro-transposable elements (SanMiguel and Bennetzen, 1998). Analysis of the genomic region surrounding the alcohol dehydrogenase gene (*Adh*) in maize has shown the presence of 23 retro-transposons in a region spanning 240 kb in maize, which are not present in the orthologous sorghum genomic region (SanMiguel and Bennetzen, 1998; Tikhonov *et al.,* 1999). These represented successive transposition events as a nested array, rather than a chaotic arrangement of retro-transposon elements. Conversely, co-linearity of functional genomic regions is usually well conserved, and conservation of non-coding regions is observed in introns or promoter elements directing protein expression (Morishige *et al.,* 2002; Ma *et al.* 2005). However, transposable elements have resulted in insertions and deletions, as well as the generation of pseudo, partial or chimeric gene copies in the maize genome (Song and Messing, 2003; Zhang *et al.,* 2006; Ilic *et al.,* 2003).

Between inbred lines there is also a high degree of polymorphism that creates complexity for gene sequence analysis (Bortiri *et al.,* 2006). A lack of gene co-linearity between inbred lines has been attributed to the action of a family of transposable elements, *Helitrons*, which do not duplicate host sequences and are hypothesized to even cause exon shuffling (Lal and Hannah, 2005). Generally allelic variations that affect both coding and non-coding sequence result in differences in protein primary structure as well as expression levels, which contribute to phenotypic diversity (Guo *et al.,* 2004).

3.1.2 Sequencing of the maize genome

Sequencing of the maize genome is near completion, made possible by technologies such as methyl-filtration and High-C_ot selection. These technologies reduced the complexity of the genome by removing repetitive elements, reducing the effective genome size six-fold (Whitelaw *et al.,* 2003). The technique of methyl-filtration is based on the theory that plant genomes are comprised of islands of genes that can be distinguished from repetitive DNA, which consists mostly of methylated retro-transposons. By removing these regions the complexity of the genome is reduced. This is done using the endogenous restriction modification system of *E. coli* that recognizes methylated DNA (Springer *et al.,* 2004). Alternatively, High-C_ot selection is based on DNA association kinetics, where concentration (C_o) and annealing time (*t*) differences result in the elimination of repetitive elements based on rates of DNA reassociation (Whitelaw *et al.,* 2003). However, large repetitive blocks of DNA are not necessarily devoid of gene sequences, and functional genes have been found dispersed within repetitive DNA regions (Tikhonov *et al.,* 1999).

The possibility that duplicated genes are mistakenly collapsed into one sequence assembly during genome sequence analysis is another source of inaccuracy. It is estimated that approximately 1% of maize genes have what is referred to as a nearly identical paralog (NIP), a percentage that is substantially higher than for other plant species such as Arabidopsis (Emrich *et al.,* 2006). However, maize genomic sequence data is available that can provide a basis for a comparative genetic map as well as information for molecular tools, such as positional cloning and gene analysis (Table 3.1).

Table 3.1. Resources for Maize Genomic Sequence Information.

Name	URL
Maize Genetics and Genomics Database	www.maizegdb.org
Maize Sequencing Project	www.maizegenome.org
Maize Assembled Genomic Island	www.plantgenomics.iastate.edu/maize
The Institute of Genomic Research	maize.tigr.org/ (www.tigr.org)

3.1.3 CA gene families

CA is a ubiquitous enzyme and is present in animals, plants, algae, archaebacteria and eubacteria. Based on primary sequence similarity, CAs can be divided into five gene families that have independently evolved the same catalytic function. The five gene families are designated alpha, beta, gamma, delta and epsilon (Hewett-Emmett and Tashian, 1996; Tripp *et al.,* 2001). Alpha-CAs in humans and animals are associated with both physiological and pathological processes resulting in extensive pharmacological research into enzyme activity and inhibition (Supuran, 2008). There is some argument that the epsilon family of CAs is merely a variant of the beta-CA class (Fabre *et al.*, 2007), and the delta-CA class was first identified in the marine diatom *Thalassiosira weissflogii* (Tripp *et al.,* 2001). The crystal structure suggests a similarity to the beta-CA dimer active site despite a lack of sequence homology (Xu *et al.,* 2008).

The gamma-CA family is thought to have evolved before the alpha-CAs (Tripp *et al.,* 2001), and until recently gamma-CAs have remained relatively uncharacterized. A gamma-CA in maize has been associated with a hydrophilic domain attached to the matrix side of Complex I of the electron transport chain in the mitochondria by crystallization (Peters *et al.,* 2008). The homologous gamma-CA of Arabidopsis has also been characterized. As part of Complex I, CA plays a role in photorespiration as well as carbon sequestration from mitochondrial catabolic reactions such as the citric acid cycle (Sunderhaus *et al.,* 2006; Parisi *et al.,* 2004; Perales *et al.*, 2005).

The beta-CAs predominate in plants and algae. In plants, multiple forms of beta-CA are common and more than one transcript encoding for a functional CA enzyme exists in many species (Burnell and Ludwig, 1997; Cavallaro *et al.,* 1994; Tetu *et al.,* 2007; Wyrich *et al.,* 1998). In *Flaveria bidentis,* a C_4 dicot species, three distinct cDNAs encoding the open reading frames of three beta-CA isoforms have been isolated. These were encoded by at least three different genes based on Southern analysis of genomic DNA (Tetu *et al.,* 2007). In sorghum, a C_4 monocot species, three transcripts encoding beta-CAs were identified by northern analysis. Sequencing of the 3′-end of the CA transcripts revealed two types of CAs that had distinct expression profiles. These two types of transcripts are encoded by two genes identified by Southern analysis of sorghum genomic DNA (Wyrich *et al.,* 1998). In contrast, rice contains only one CA transcript and one CA gene was identified based on Southern hybridization with the transcript sequence (Wyrich *et al.,* 1998).

The CA transcripts from C_4 monocot species are distinct from dicot CAs based on several unique characteristics (Burnell, 2000). In maize, three transcripts encoding beta-CAs have been identified and designated CA1, CA2 and CA3 (Fig. 3.1; Burnell and Ludwig, 1997). The third and smallest transcript (1.3 kb) has been designated CA1 (Burnell, unpublished), while the longest transcript (gi:606814) is referred to as CA3. The number of CA genes in maize remains to be determined.

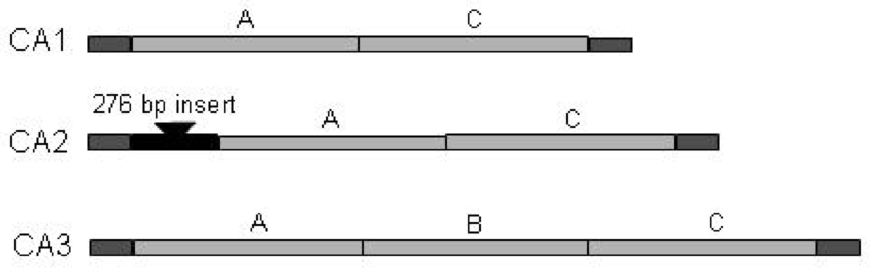


Fig. 3.1. *Schematic representation of the three beta-CA cDNAs from maize (Burnell and Ludwig, 1997; Burnell, unpublished). The repeat sequences are labelled A, B and C and are approximately 600 bp. The black shaded box represents the 276 bp insert unique to CA2. The dark grey shaded boxes represent the 5′-leader sequences and 3′-untranslated regions.*

3.1.4 Rationale

The characterization of the CA gene family in maize follows from the early discovery of more than one CA isozyme in many plant species (Atkins *et al.*, 1972a), and specifically the identification of three CA mRNA transcripts in maize (Burnell and Ludwig, 1997). There was 7.5 kb of CA genomic sequence, but this was not a complete gene and represented only the first nine exons, eight of which encoded the first protein domain (Repeat A; Burnell, 2000). The last exon in the sequence encoded the start of the second protein domain. The second and longest exon in the CA gene represented the 276 bp insert that is unique to the CA2 cDNA sequence, and hence the 7.5 kb genomic clone corresponded to the gene encoding CA2 (Fig. 3.1).

To enable isolation of the full-length CA2 gene a maize genomic DNA library was screened, and to determine if this was the only gene encoding CA in the maize genome, Southern blot analysis of leaf-extracted genomic DNA was performed. Interrogation of maize genomic databases (Table 3.1) provided sequence data enabling investigation of gene structure and orthologous genes in related species. In no single plant species has the total number of CA genes, and the location and function of the isozymes they encode, been fully elucidated.

Materials and methods

3.2.1 Purification of genomic DNA

Genomic DNA was purified from the leaves of maize plants, either fresh or from leaf material stored at -80ºC (Section 2.3, Chapter 2). Leaves were ground thoroughly in a mortar and pestle with liquid nitrogen and up to 5 $ml.g^{-1}$ of extraction buffer (0.1 M Tris-HCl pH 9.0, 0.1 M NaCl, 1% w/v SDS, 50 mM DTT).

An equal volume of phenol/chloroform (1:1) was then added to the extraction mixture and incubated at 37ºC for 30 min or at room temperature for 16 h. The mixture was transferred to 15 ml Corex (glass) centrifuge tubes and centrifuged for 15 min at 1,000 × *g*. The top aqueous phase was removed and re-extracted with phenol/chloroform/isoamylalcohol (25:24:1) twice.

In a 15 ml Corex tube, 0.1x vol of cold 3 M Na acetate pH 5.2 was added to the top aqueous layer and mixed briefly before the addition of 2.5x vol of cold 95% ethanol. The Corex tube was then gently rotated until strands of genomic DNA were visible and these were removed by spooling onto a glass pipette and transferred into a 1.5 ml microcentrifuge tube. The genomic DNA was washed several times with cold 70% ethanol before being resuspended in TE buffer.

The genomic DNA was treated with RNase A (Sigma), which was prepared by dissolving in distilled water, boiled for 30 min and then slowly cooled to room temperature (stored at -20ºC). RNase A was added to the genomic DNA to a final concentration of 0.05 $\mu g.ml^{-1}$ and incubated at 65ºC for 10 min, before the addition of 0.6x vol 5 M ammonium acetate and further incubation at 4ºC for 30 min. After centrifugation at 16,100 × *g* for 30 min at 4ºC, the genomic DNA in the supernatant was precipitated by the addition of 2.5x vol cold 95% ethanol at -20ºC for 60 min. The genomic DNA was pelleted by centrifugation at 16,100 × *g* for 20 min at 4ºC and the pellet resuspended in TE buffer. The DNA concentration and purity were assessed by measuring the absorbance at 260 nm, and the $A_{260/280}$ ratio (Section 2.1.5, Chapter 2).

3.2.2 Analysis of gene number by Southern blotting

The method used was adapted from that previously described (Sambrook *et al.,* 1989).

3.2.3 Restriction endonuclease treatment and electrophoresis of genomic DNA

Restriction enzymes (600 units) were added directly to 100 µg of genomic DNA in a 1.2 ml total reaction volume. The appropriate 10x reaction buffer and BSA (where required to a final concentration of 100 µg.ml^{-1}), were added and the reactions incubated for 48 h at 37ºC. After 24 h, the reactions were spiked with an additional 600 units of enzyme.

Genomic DNA that had been treated with restriction enzymes was precipitated by the addition of 0.1x vol cold 3 M Na acetate, pH 5.2 and 2.5x vol cold 95% ethanol and incubated at -20ºC for at least 60 min. The DNA was pelleted by centrifugation at 16,100 × *g* for 20 min at 4ºC. The pellet was washed with cold 70% ethanol and resuspended in a minimal volume of TE buffer.

The DNA fragments (10 µg) generated by restriction endonuclease treatment were resolved on a 0.8% (w/v) agarose gel (Section 2.1.2, Chapter 2), dimensions 25 cm x 14 cm. Electrophoresis was performed at 30 V for 17 h, with migration monitored by comparison to the mobility of molecular weight markers. DNA was visualized after ethidium bromide staining using a UV transilluminator.

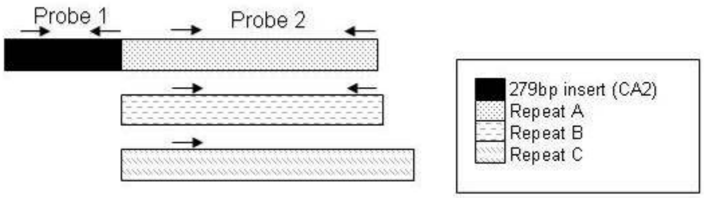
3.2.4 Southern blotting to nylon membrane

The DNA in the gel was depurinated by incubation of the gel in 0.25 M HCl for 15 min, followed by two 20 min washes in denaturing buffer (0.5 M NaOH, 1.5 M NaCl). The gel was briefly rinsed with distilled water then washed in neutralization solution (1 M Tris-HCl pH 8, 1.5 M NaCl) twice for 20 min. The transfer was performed in 0.4 M NaOH or 20x SSC (3 M NaCl, 0.3 M Na citrate, pH 7.0) to Hybond N+ (Amersham Biosciences) positively charged nylon membrane.

3.2.5 Preparation of probes

The two probes used for Southern blotting were CA cDNA fragments generated by PCR using CA2 cDNA template (Section 2.1.14, Chapter 2). The first probe represented one exon unique to the CA2 sequence, and the second probe contained sequence covering several exons that corresponded to Repeat A of the CA2 cDNA sequence (Genbank accession U08401, version gi:606810; Fig. 3.2). The forward primer used to generate this probe bound also to Repeat B and C, while the reverse primer had homology to Repeat B as well as Repeat A. As CA2 only contained Repeat A and Repeat C, only Repeat A was amplified from this primer combination.

(A)



(B)

	Probe 1 - *Insert*	Probe 2 - *Repeat*
Forward	5′-ggcgggcataagagggg-3′	5′-gtccatggtgttcgcctgctcc-3′
Reverse	5′-gcccttggaggaagccttggaggg-3′	5′-cctagggaggctcccatgtgacg-3′

(C)

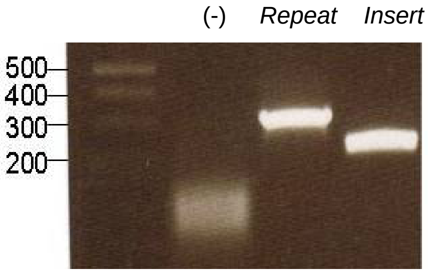


Fig. 3.2. *(A) Schematic representation of the method for amplifying the probes used for hybridization with maize genomic DNA. (B) The primer combinations used to amplify the probes used. (C) The PCR products were visualized by agarose gel electrophoresis and ethidium bromide staining (see Appendix 3.1 for probe sequences).*

The DNA fragment was visualized by ethidium bromide staining of an agarose gel, excised using a scalpel and extracted using a QIAGEN gel extraction kit and protocol (Section 2.1.3, Chapter 2). Labelled probe was prepared according to the protocol of the DECAprime II Random Primed DNA Labelling Kit (Ambion, USA), using radioactive isotope ^{32}P (PerkinElmer Life and Analytical Sciences Inc, USA). Unincorporated [α-^{32}P]-dCTP was removed by chromatography on a Sephadex G-50 (Fine) column. The disintegrations per minute (dpm) of the labelled probe fragment were counted using a BioScan radioisotope counter QC:4000 XER (BioScan, USA).

3.2.6 Hybridization and detection

The DNA fragments were cross-linked to the membrane by either incubation at 80ºC for 2 h, or by exposure to UV illumination for 5 min. Where necessary, the membrane was stored dry at room temperature before use. The membrane was blocked in pre-hybridization solution (0.1% v/v SDS, 5x Denhardt's solution, 4x SSC [0.15 M sodium chloride, 0.015 M sodium citrate], 0.05 M inorganic phosphate, 50% v/v formamide and 100 µg.ml^{-1} denatured salmon sperm DNA) in a 42ºC incubator for at least 1 h before addition of the labelled probe. Alternatively, the membrane was pre-hybridized in 10 ml of 50% NEN hybridization mix (50% v/v deionised formamide, 1 M NaCl, 1% w/v SDS and 10% w/v dextran sulphate) at 42°C for 1 h.

Hybridization was conducted at 42°C for 16 h using 1 x 10^{6} dpm of denatured ^{32}P-labelled probe fragments per 10 ml of 50% NEN hybridization mix. The hybridized membranes were washed twice in 2x SSC buffer containing 1% (w/v) SDS for 30 min at 65ºC and once in 0.2x SSC containing 1% (w/v) SDS for 30 min at 65ºC with agitation.

The membranes were exposed by placing in a Phosphorimager cassette (Molecular Dynamics) at -80ºC for up to one week. The image produced was analysed using ImageQuant software. Alternatively the membrane was exposed to Kodax BioMax MS X-Ray film (Kodak, USA) using intensifying screens overnight at -80ºC and the film was automatically developed using a Curix 60 X-ray developer (AGFR-Gevaert Group, Belgium).

In order to reprobe the same membrane with a different radio-labelled probe, hybridized ^{32}P-labelled fragments of probes were removed from Southern blots in two washes of boiling 0.1% (w/v) SDS for 30 min at 65°C. The 0.1% (w/v) SDS solution was then allowed to cool to room temperature and the membrane rinsed in 2x SSC prior to being exposed to Kodax BioMax MS X-Ray film (Kodak, USA) with an intensifying screen at -80°C for 16 h to ensure adequate removal of the probe.

3.2.7 Calculation of the molecular weight of hybridizing fragments

A standard ruler was used to determine the distance migrated for both the molecular weight markers used and the hybridizing bands. A standard curve was generated using the molecular weight marker data and this was used to determine the molecular weight of the hybridizing bands.

3.2.8 Screening the maize genomic DNA library

The maize genomic DNA library used for this analysis was prepared from etiolated seedlings of maize plants (Missouri 17 inbred line, Stratagene, Vector: Lambda FIX® II vector, Insert Size: 9 - 23 kb).

3.2.9 Preparation of host cells

XL-1 Blue MRA P2 cells were streaked from a glycerol stock onto an LB agar plate and grown overnight at 37ºC. A single colony was used to inoculate 10 ml of LB in a 50 ml Falcon tube and shaken at 37ºC until the late log phase of growth was reached. This culture was used to make glycerol stocks, which were stored at -80ºC.

An LB agar plate was streaked from a prepared glycerol stock and grown overnight at 37ºC. A single colony was used to inoculate 50 ml of LB in a 250 ml flask that had been supplemented with 0.2% (w/v) maltose and 10 mM $MgSO_4$ and was shaken overnight at 30ºC. The cells were harvested by centrifugation at 750 × *g* for 10 min and the pellet resuspended in 10 mM $MgSO_4$ so that the optical density at 600 nm was 0.5.

3.2.10 Genomic DNA library titration and plating phage

The genomic DNA library was titred in order to obtain 50,000 plaque forming units (*pfu*) per plate (150 mm diameter). Serial dilutions (10^{-2} to 10^{-5}) of the genomic DNA library were created in SM buffer (50 mM Tris-HCl, pH 8, 100 mM NaCl, 8 mM $MgSO_4$, 0.01% w/v gelatin), 1 µl of which was added to 500 µl of host cells and incubated at 37ºC for 20 min.

To this, 10 ml of 0.75% (w/v) LB top agar was added and poured onto large pre-warmed LB agar plates and incubated overnight at 37ºC. The number of *pfu* were then determined.

3.2.11 Genomic DNA library screening

For the purposes of screening the genomic DNA library, 2.5 µl of 10^{-3} library dilution was used to pour six large LB agar plates. These were incubated for approximately 8 h at 37ºC. Hybond N+ (Amersham Biosciences) positively charged nylon membranes were used to make plate lifts. The membranes were placed over the plates for approximately 3 min, transferred to denaturing solution-saturated blotting paper for 2 min, then to neutralization solution-saturated blotting paper for 2 min. The membranes were rinsed briefly with 2x SSC (0.3 M NaCl, 30 mM Na citrate, pH 7.0) before cross-linking at 80ºC for 2 h.

The probe used for the genomic DNA library screen was a 900 bp fragment generated from restriction endonuclease treatment (*Eco*RI/*Xho*I) of a plasmid construct containing 1,638 bp of CA2 cDNA sequence (Genbank accession number U08401, version gi:606810, Fig. 3.3). This fragment corresponded to the 5′-end of the CA2 sequence including the 5′-leader sequence, the unique 276 bp insert and Repeat A. Probe labelling, hybridization and detection were performed as described in Section 3.2.2.3 and 3.2.2.4.

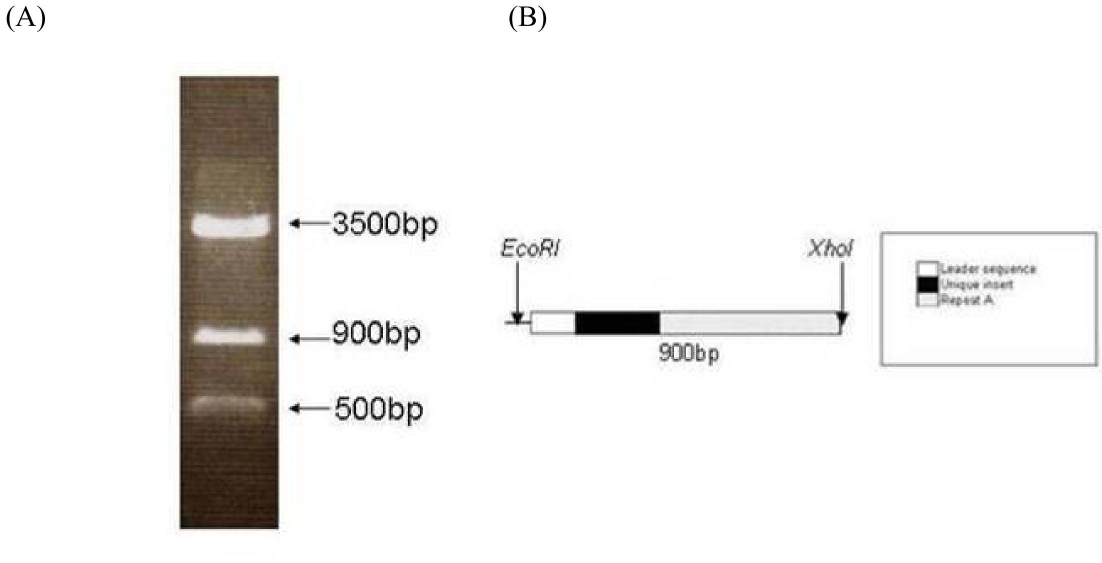


Fig. 3.3. *(A) Agarose gel electrophoresis of the* Eco*RI*/Xho*I treated CA2 plasmid construct. The 900 bp band was excised from the gel, purified and used to probe the maize genomic DNA library. (B) Schematic representation of the components of the CA2 probe sequence. The* Eco*RI recognition site was before the 5'-end of the cDNA sequence, in the multiple cloning site of the vector (see Appendix 3.2 for probe sequence).*

Positive plaques were identified by aligning with detected spots, and removed using a cut pipette tip into 200 µl of SM buffer with 5 µl of chloroform and allowed to elute at 4ºC overnight. Secondary screens were performed by re-plating these excised plaques as necessary.

3.2.12 Preparation of lambda DNA

The positive plaque was amplified by incubating 1 µl of the excised plaque with 600 µl host cells for 20 min at 37ºC, before plating onto six large LB agar plates using 8 ml 0.75% (w/v) LB top agar. Following overnight incubation at 37ºC, 6 ml SM buffer was poured over each plate and incubated at 4ºC with gentle agitation for up to 6 h. The top agar was scraped from each plate into 50 ml Falcon tubes, 500 µl of chloroform was added, and the cells lysed by shaking at room temperature for 30 min. The cellular debris was pelleted by centrifugation at 2,900 × *g* for 15 min and 0.25x volume of PEG reagent (5x: 207 g PEG

6000/8000, 6 g dextran sulphate, 49.5 g NaCl per 350 ml distilled water) was added to the supernatant, which was gently mixed on ice for 1 h. The pellet (a yellow disc) was obtained by centrifugation for 20 min at 2,900 × *g*. The pellet was gently resuspended in 2 ml SM buffer and 650 µl of 4 M KCl was added; the solution was mixed thoroughly and centrifuged for 10 min at 2,900 × *g*.

The lambda DNA was purified using a CsCl gradient. The gradient was made by layering 3.5 ml of 1.7 g.ml^{-1}, 2.5 ml of 1.5 g.ml^{-1} and 2.5 ml of 1.3 g.ml^{-1} CsCl in Beckman centrifuge tubes (14 mm x 89 mm). The gradient was centrifuged at 35,000 rpm (Beckman SW41 rotor) for 60 min at 4ºC. The bluish band was carefully extracted through the side of the tube using a needle and syringe. To this, 0.1x vol of 1 M Tris-HCl pH 8, 0.04x vol of 0.5 M EDTA and 1x vol formamide were added and the solution incubated at room temperature for 1 h. To precipitate the DNA, 1x vol of distilled water and 6x vol of cold 95% ethanol were added and the tube swirled until the DNA was visible (alternatively the solution was placed at -20ºC overnight). The DNA was removed by winding around a glass rod and transferred to a microcentrifuge tube, washed several times with cold 70% ethanol and the pellet resuspended in 500 µl TE buffer.

3.2.13 Restriction endonuclease treatment of genomic DNA library clone

Purified phage DNA was treated with restriction endonucleases in the appropriate buffer at 37ºC for 3 h (Section 2.1.1, Chapter 2). DNA fragments were separated by agarose gel electrophoresis, visualized by ethidium bromide staining and fragments of interest were excised from the gel and purified for ligation with vector (pGEM®-T, Promega; Section 2.1.11, Chapter 2).

3.2.14 Southern hybridization of lambda DNA

Further analysis of the genomic DNA library was performed by Southern blotting the sub-cloned DNA fragments that had been generated by restriction endonuclease treatment of the purified lambda DNA clone isolated during the genomic DNA library screen (Sections 3.2.2.2-3.2.2.5).

3.2.15 Sequence analysis

Sequences were aligned and translated using MacVector^{TM} 7.0 and ClustalW sequence analysis software (Thompson *et al.,* 1994). The sequence of AY109272 and DQ246083 were obtained from the NCBI database (http://www.ncbi.nlm.nih.gov).

3.2.16 Amplification of CA genomic DNA

PCR was performed using primers designed to bind conserved exon sequence, while spanning potential intron DNA. The primers used were 5′-ggcgggcataagagggg-3′ and 5′-gcccttggaggaagccttggaggg-3′ (Fig. 3.4, Table 3.2). The template for these reactions was genomic DNA extracted from maize leaf tissue (Section 3.2.1). Products generated were sub-cloned and the sequence analyzed.

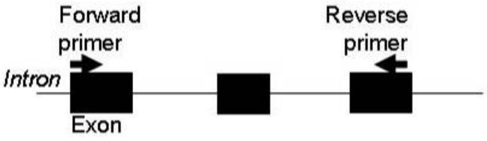


Fig. 3.4. *Schematic representation of the primer binding sites used to amplify CA genomic DNA and spanning both exons and introns.*

Table 3.2. Reaction conditions for amplification of CA genomic DNA.

Cycling Conditions	Step	Temperature	Time
	1	95ºC	5 min
	2	95ºC	1 min
	3	51ºC	1 min
30 cycles from Step 2	4	72ºC	3 min
Final extension	5	72ºC	10 min

3.3 Results

3.3.1 The maize CA2 gene

3.3.1.1 Sequence of the maize CA2 gene

Approximately 7.5 kb of a CA genomic DNA clone sequence was available, which corresponded to the CA2 gene (Burnell, 2000). This sequence included 1,500 bp of genomic DNA upstream of the first exon, one exon that corresponded to the unique 276 bp insert of CA2, the exons composing the first repeating region of the CA cDNA sequences (Repeat A), but containing only the first exon of Repeat C (Fig. 3.5). To obtain the full length sequence of the CA2 gene, a maize genomic DNA library was screened.

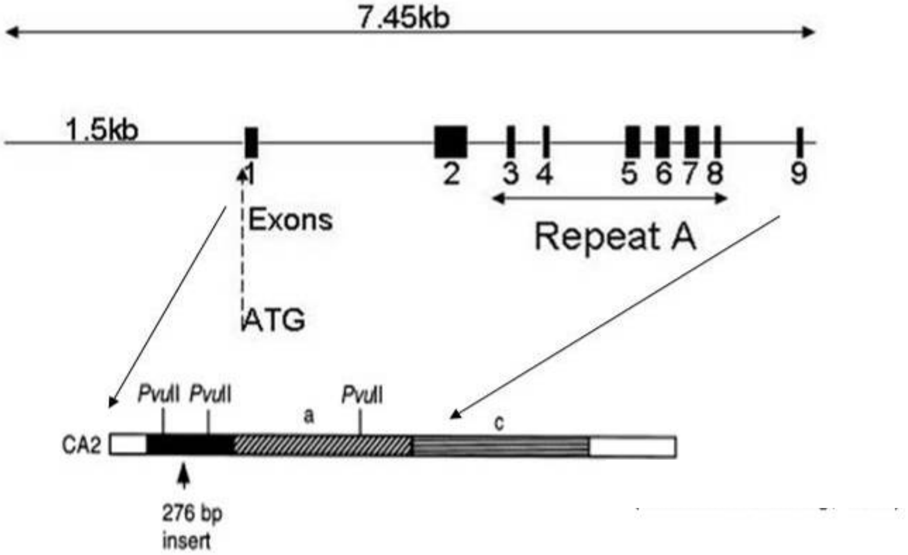


Fig. 3.5. Schematic representation of the genomic 7.45 kb sequence that corresponded to the CA2 gene and contained the first nine exons of the CA2 cDNA sequence (Genbank accession U08401, version: gi:606810; Burnell and Ludwig, 1997).

3.3.1.2 Screening the maize genomic DNA library

A maize genomic library (Lambda Fix II, Stratagene) was screened using a 900 bp probe excised from a CA2 plasmid construct using *Eco*RI and *Xho*I (Appendix 3.2). This fragment corresponded to the cDNA sequence of the 5′-end of CA2 including the 5′-leader sequence, the unique 276 bp insert and Repeat A.

3.3.1.3 Analysis of the genomic DNA library clone

A positive clone was identified and isolated from the library screen. The clone was treated with a number of restriction enzymes to generate smaller fragments of genomic DNA, which were sub-cloned and sequenced. Southern analysis using the same probe as that used to screen the library enabled identification of restriction fragments that contained sequence of interest (Fig. 3.6). These were excised, and sub-cloned using the pGEM®-T vector system (Promega).

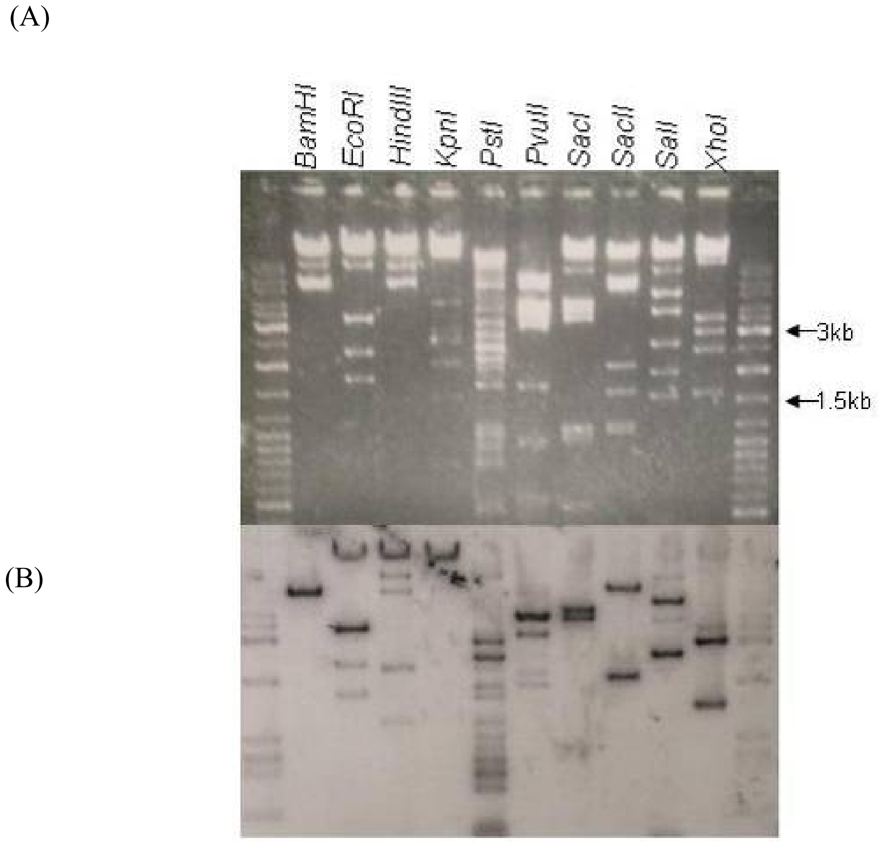


Fig. 3.6. *(A) Agarose gel electrophoresis of the genomic library clone after restriction endonuclease treatment with the restriction enzymes indicated. The molecular weight markers are in lanes 1 and 12. (B) Probing the restriction fragments with the 900 bp probe used to screen the genomic DNA library enabled identification of bands that contained sequence of interest, which could then be sub-cloned for further analysis.*

After sub-cloning the DNA fragments that were identified to contain CA sequence, these constructs were analyzed and approximately 3,500 bp of sequence was assembled. This sequence was homologous with the CA2 gene (Appendix 3.3). Additionally the 3′-end of the genomic sequence was isolated from the library screen and approximately 2,700 bp was assembled and included the last three exons corresponding to the CA2 cDNA sequence (Fig. 3.7; Appendix 3.3).

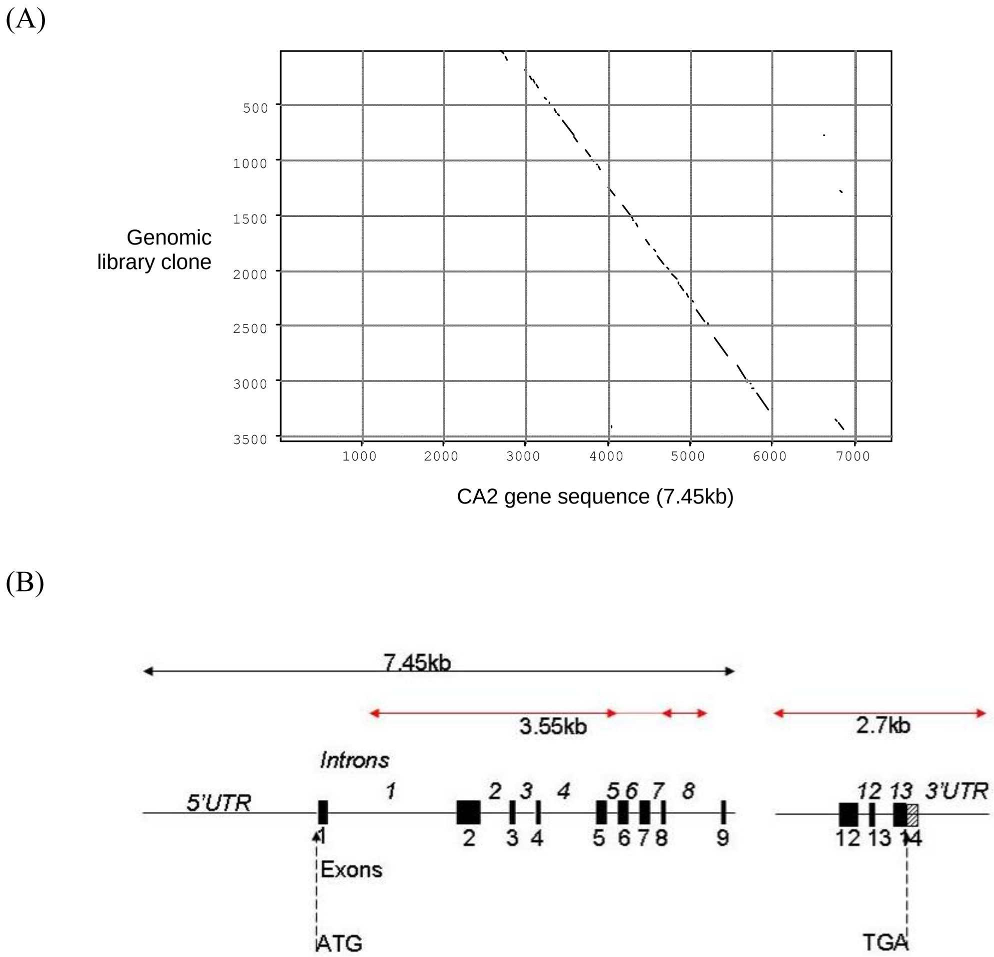


Fig. 3.7. *(A) Matrix alignment plot showing similarity between the CA2 gene sequence and the genomic library clone assembly. Similarity is based on a minimum alignment of 90% in a 30 bp window (hash value 6). (B) Schematic representation of the alignment of the assembled genomic library sequence, red, with the CA2 gene sequence.*

3.3.2 Beta-CA multi-gene family

3.3.2.1 Determining gene number by Southern analysis

Southern blotting using genomic DNA allows identification of gene size and copy number, where more than one gene containing homologous sequence is present in an organism. Maize genomic DNA that had been purified from leaf tissue and modified with restriction enzymes was analysed by Southern blotting and hybridization with randomly ^{32}P-labelled maize CA2 cDNA sequence.

The purified genomic DNA was treated with restriction endonucleases, based on the expected frequency of recognition sites allowing prediction of the number of hybridizing bands. With these predications, a comparison could also be made to the results obtained for the maize genomic DNA library screen (Fig. 3.6). The two probes used for Southern blotting were CA cDNA fragments generated by PCR using CA2 cDNA template. The first probe represented one exon unique to the CA2 sequence, and the second probe contained sequence covering several exons that correspond to Repeat A of the CA2 cDNA sequence (Fig. 3.2). The sequence of the probes was confirmed before use.

After hybridization with both probes, the hybridization band patterns obtained gave an indication of the number of CA genes in the maize genome (Fig. 3.8).

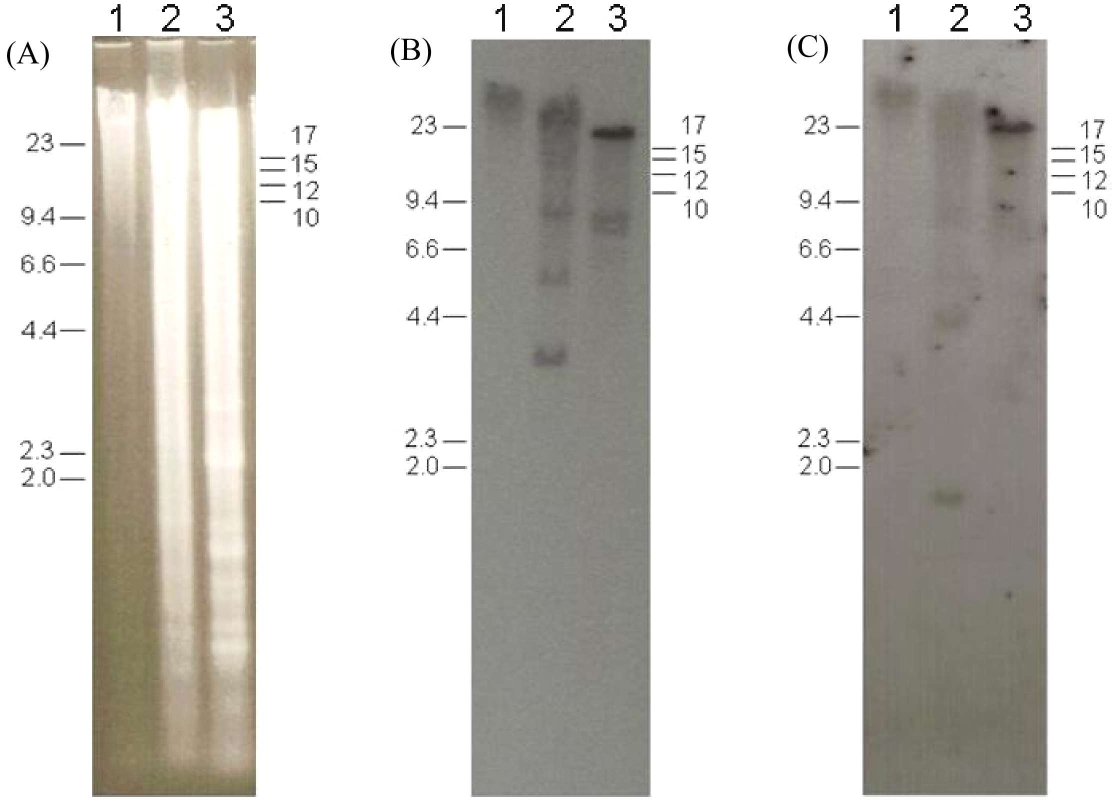


Fig. 3.8. *(A) Agaroge (0.8%) gel electrophoresis of 10 µg of maize genomic DNA; Southern blot analysis of maize genomic DNA (10 µg) (B) probed with the Repeat probe, and (C) probed with the Insert probe. (1) untreated, (2) treated with* Eco*RI and (3) treated with* Kpn*I. The molecular weight markers used are Hyperladder VI (Bioline) and λ*Hind*III (Promega).*

The molecular sizes of the observed hybridizing band were determined by the generation of a standard curve (Fig. 3.9, Table 3.3).

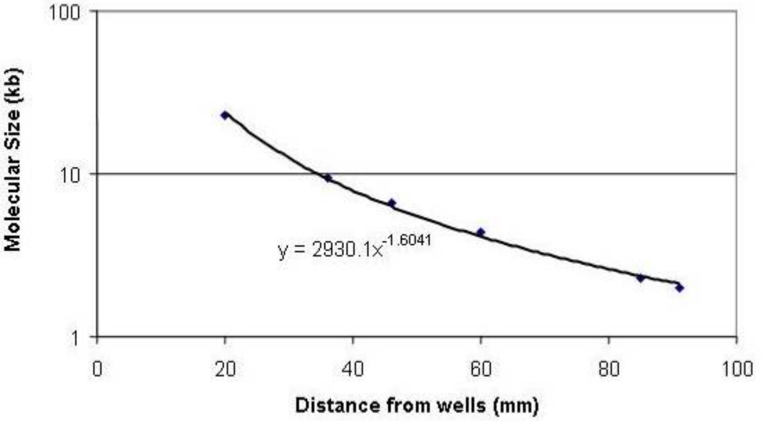


Fig. 3.9. *Standard curve relating DNA fragment size (molecular weight markers) and distance migrated (mm).*

Table 3.3. The estimated molecular weight of hybridizing bands observed (Fig. 3.7), determined using the standard curve generated (Fig. 3.8).

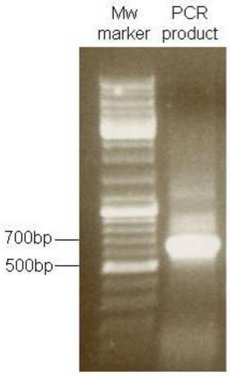
Probe	*Insert*				*Repeat*			
			Distance migrated (mm)	Size (kb)			Distance migrated (mm)	Size (kb)
	*Eco*RI	1	40	7.89	*EcoRI*	1	37	8.94
		2	52	5.18		2	51	5.34
		3	60	4.12		3	68	3.37
		4	97	1.91				
	*Kpn*I	1	22	20.58	*Kpn*I	1	22	20.58
						2	39	8.22
						3	41	7.58

3.3.2.2 PCR analysis of maize genomic DNA

PCR was performed using primers designed to bind conserved exon sequence, while spanning potential intron DNA (Section 3.2.5). Products generated were sub-cloned and the sequence analyzed, which provided both a basis for comparison to known sequence and the identification of novel CA genomic sequence.

The reaction with a primer combination that hybridized within the repeating portion of the CA cDNA sequences produced one predominant product when analysed by agarose gel electrophoresis, however consisted of two DNA fragments of similar lengths (Fig. 3.10). When the sequence of these products was analysed, it was found that the exon sequence and intron length were conserved, while the intron sequences were different, indicating the presence of at least two genes encoding CA in the maize genome.

(A)



(B)

Exon 5

```
Product 1 GTACATGGTGTTCGCCTGCTCCGACTCCCGCGTGTGCCCGTCGGTGACACTGGGCCTGCAGC
Product 2 GTACATGGTGTTCGCCTGCTCCGACTCCCGCGTGTGCCCGTCGGTGACCCTGGGCCTGCACC
          ************************************************ *********** *

Product 1 CCGGCGAGGCATTCACCGTCCGCAACATCGCCTCCATGGTCCCACCCTACGACAAG
Product 2 CCGGCGAGGCCTTTGCCGTCCGCAACATCGCCAGCATGGTGCCGCCCTACGACAAG
          ********** **  *****************  ****** ** ************
```

92.8%

Intron 6

```
Product 1 GTACGTACGTACGAGCAAACACCGATCGACGCATGCAACGG-TGGTATCAGCCACACTAA
Product 2 GTGAGCACACGCGCGCACGACGCATGCATCGTACGCCTCCCCTGGTAACAACTGTGTGTG
          **    **   ** ***     *   *  **   **  *   ***** ** *

Product 1 TATTACTCACACGGTCGTCTTCCGTTTTGGCCAAACTGCAG
Product 2 GCCTCTAGCGACTCACGCGTACTATTGTCGATCGACTGCAG
             *      **   **  * *  ** * *    *******
```

41.7%

Fig. 3.10. *(A) Agarose gel electrophoresis of the products of the reaction used to amplify CA genomic DNA including both exon and intron gene components. (B) The sequences of the DNA amplified, specifically the first exon and first intron amplified.*

3.3.2.3 Interrogation of the Maize Genomic Databases

The sequence data obtained from PCR analysis (Section 3.3.2.2) was used to search genomic maize databases for CA genomic sequence. The databases used for the BLAST (basic local alignment search tool) analysis included the Maize Genetics and Genomics Database (www.maize.gdb.org), the Maize Sequencing Project (www.maizegenome.org), the Institute of Genomic Research (www.tigr.org) and the Maize Assembled Genomic Island (http://magi.plantgenomics.iastate.edu).

There were at least four sequences identified from the databases that appeared to encode CA in the maize genome. These were:

- AZM4_68974
- AZM4_68973
- AZM4_38976
- AZM4_23203

The first assembly, AZM4_68974, represented genomic sequence corresponding to the 5′-end of the CA2 cDNA sequence. It was homologous with the CA2 gene and Clone 1 obtained from the genomic DNA library screen (Fig. 3.11; Appendix 3.4). The AZM4_68974 sequence extended a further 1,000 bp upstream from the 5′-end of the CA2 gene sequence.

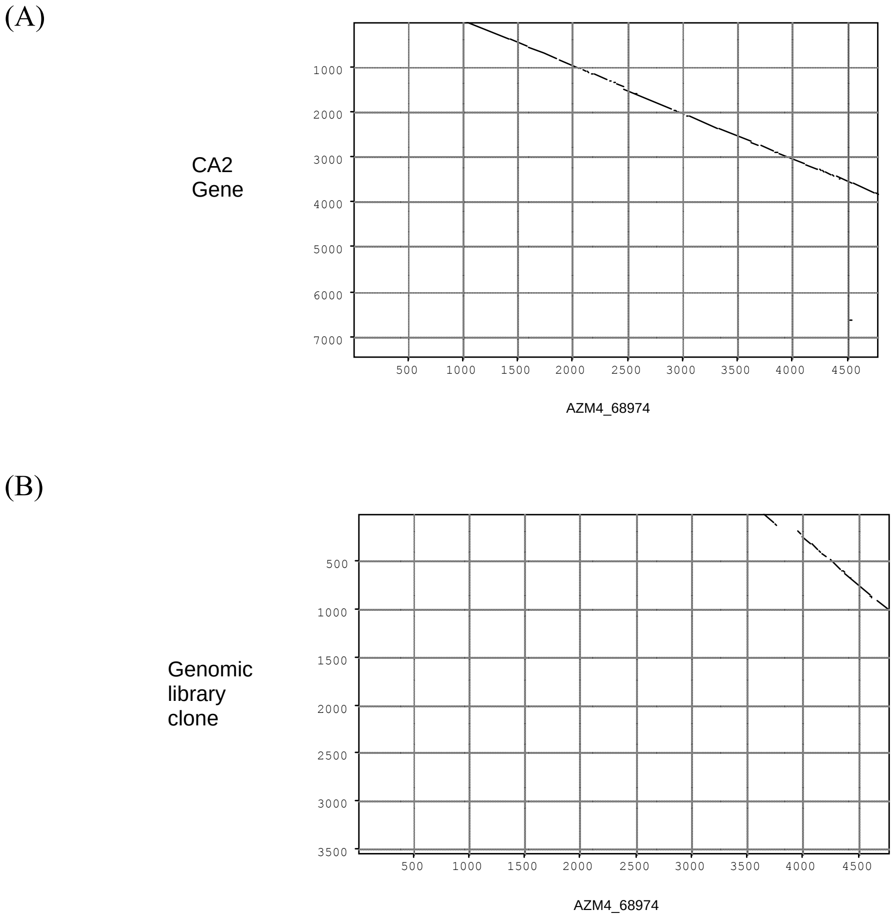


Fig. 3.11. *Matrix alignment plot showing (A) similarity between the AZM4_68974 assembly and the previously sequenced 5′-end of the maize CA2 gene, and (B) similarity between the AZM4_68974 assembly and the genomic library clone. Similarity is based on a minimum alignment of 80% in a 20 bp window (hash value 6).*

The AZM4_68974 assembly corresponded to 2,500 bp upstream of the first exon, the first exon (5′-leader sequence), and the 276 bp insert unique to CA2 (Fig. 3.12).

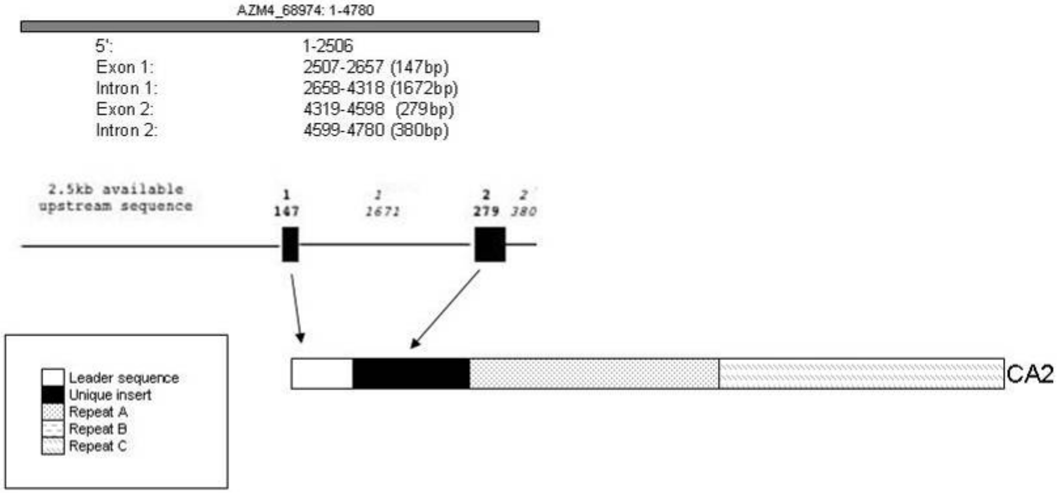


Fig. 3.12. *Schematic representation of the AZM4_68974 assembly, with exons indicated that corresponded to the CA2 cDNA sequence.*

The second assembly AZM4_68973 included the remaining exons of the CA2 cDNA sequence and also had homology with the 3′-end of the CA2 gene sequence and the genomic DNA library clone (Fig. 3.13). AZM4_68974 and AZM4_68973 represented two adjacent assemblies that corresponded to the full length CA2 gene.

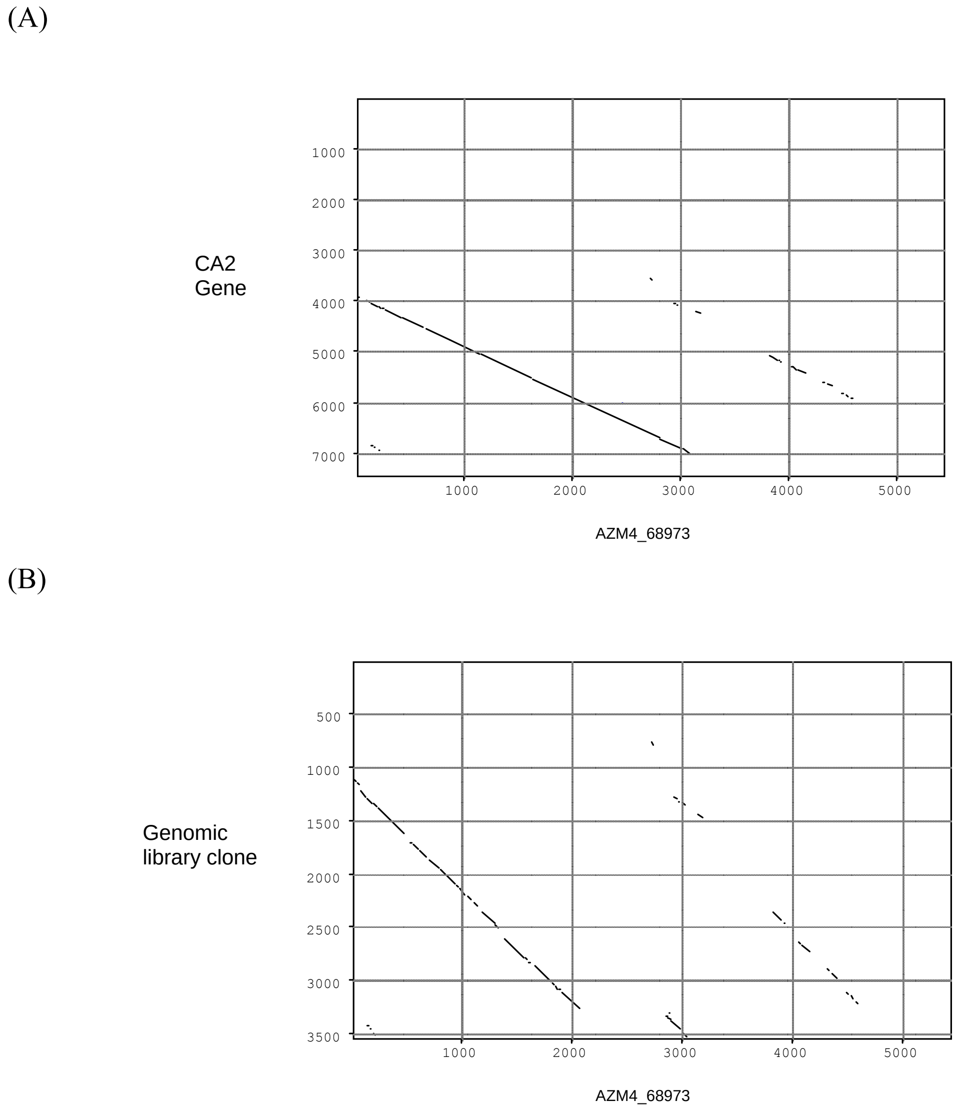


Fig. 3.13. *Matrix alignment plot showing (A) similarity between the AZM4_68973 assembly and the CA2 gene sequence, and (B) similarity between the AZM4_68973 assembly and the genomic DNA library clone. Similarity is based on a minimum alignment of 80% in a 20 bp window (hash value 6).*

AZM4_68973 included the exons composing Repeat A and Repeat C of CA2 and included approximately 560 bp of sequence after the 3′-untranslated region (Fig. 3.14; Appendix 3.5 and 3.6).

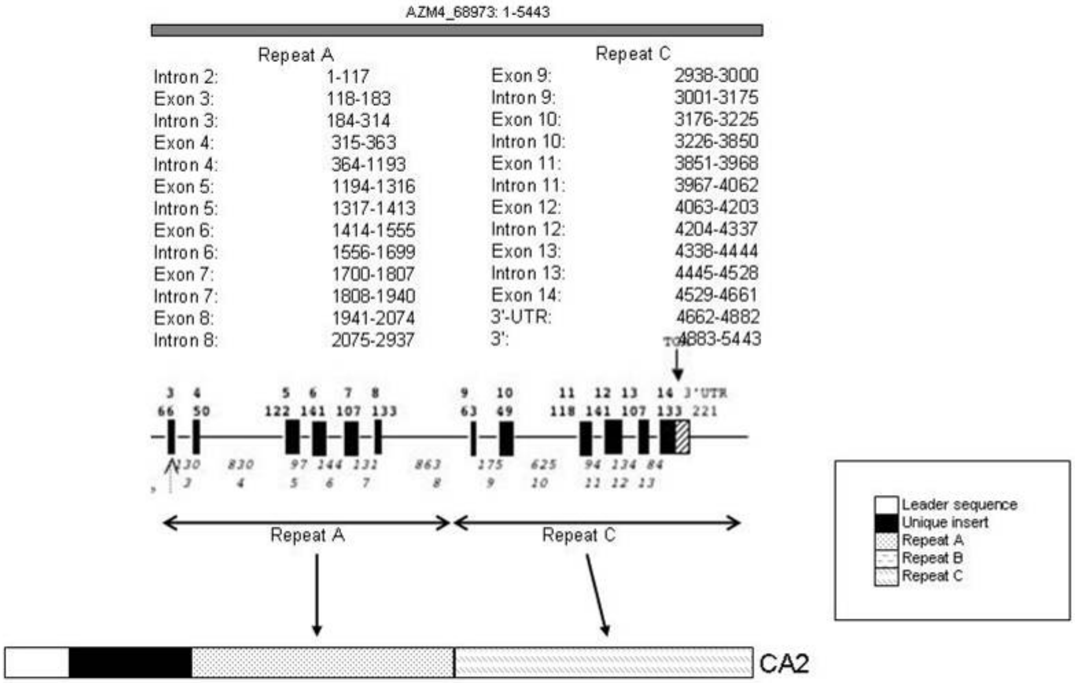


Fig. 3.14. *Schematic representation of the AZM4_68973 genomic sequence, with exons indicated that corresponded to the CA2 cDNA sequence.*

The other two assemblies, AZM4_38976 and AZM4_23203 had different intron sequences compared to AZM4_68973, however the exon sequence was conserved. The AZM4_38976 assembly was 3,876 bp and had homology with CA cDNA sequences when subjected to a BLASTn homology search (NCBI database), however unlike the AZM4_68973 assembly an exact match with available cDNA sequence (for example, CA2 and AZM4_68973/4) was not determined. As well as CA2, there are two cDNA sequences available on NCBI (AY109272 and DQ246083) that are apparent CA transcripts. Genbank entry AY109272 (gi:21212748) appeared to be the mRNA product of the AZM4_23203 assembly (Fig. 3.15; Appendix 3.7).

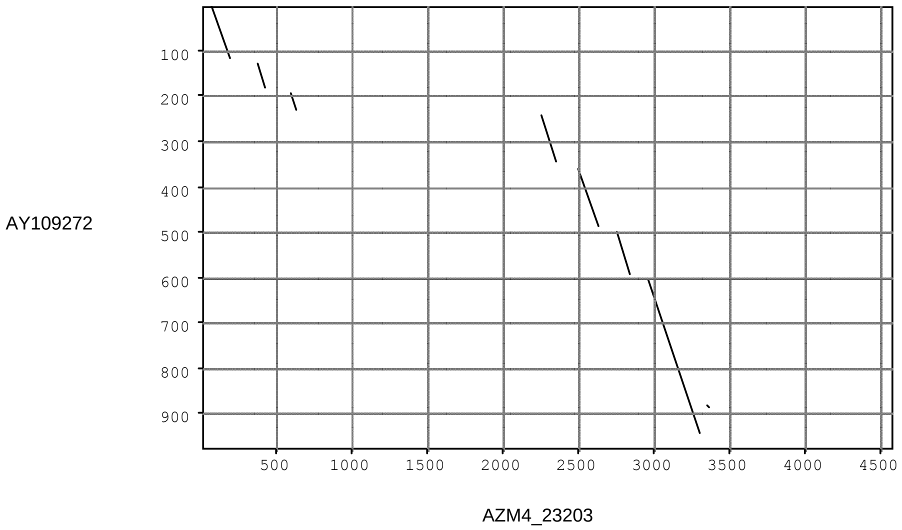


Fig. 3.15. *Matrix alignment plot showing similarity between the AZM4_23203 assembly and the AY109272 mRNA sequence, indicating the exon/intron structure. Similarity is based on a minimum alignment of 80% in a 20 bp window (hash value 6).*

The AZM4_23203 assembly included the seven exons that compose the mRNA represented by AY109272 (Fig. 3.16). Unlike the CA2 gene, there was not a large intron between the first two exons; rather a relatively large intron existed between the third and fourth exons. Additionally, the last exon of AY109272 was relatively long, almost 300 bp.

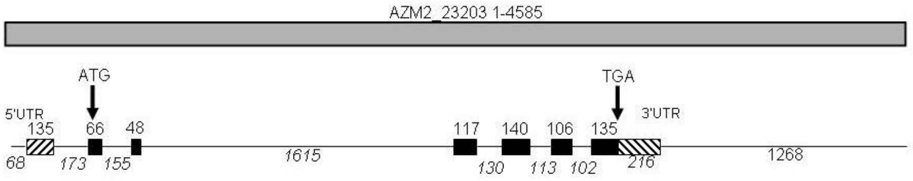


Fig. 3.16. *Schematic representation of the AZM2_23203 genomic sequence.*

The AY109272 and DQ246083 translated sequences were 91% homologous, and were also 86% and 83% homologous with the translated sequence of Repeat A respectively (Fig. 3.17).

```
AY109272        MGDAVEHLKSGFQKFKTEVYDKKPELFEPLKAGQAPKYMVFACSDSRVCPSVLG-LQPGE
DQ246083        MDDPVERLKDGFHKFKTEVYDKKPELFEPLKAGQAPKYMVFACSDSRVCPSVTLGLQPGE
Repeat_A        MDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVCPSVTLGLQPGE
                *. .**:**.**:******************:**:*:***************   *****

AY109272        AFTVRNIAAMVPAYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIRALLSLQDGAP
DQ246083        AFTVRNIAAMVPAYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIRALLSLQDGAP
Repeat_A        AFTVRNIASMVPPYDKIKYAGTGSAIEYAVCALKVQVIVVIGHSCCGGIRALLSLKDGAP
                ********:***.*** **:* *************:*:*****************:****

AY109272        DNFHFVENWVKIGFPAKVKVKKEHASVPFDDQCSILEKEAVNVSLENLKTYPFVQEGLAK
DQ246083        DTFHFVENWVKIGFPAKIKVKKDHASVPFDDQCSILEKEAVNLSLENLKTYPFVKDGLAN
Repeat_A        DNFHFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEGLAG
                *.*****:**:** *** ****:*******************:**:***:****::***

AY109272        GTLKLVGAHYDFVSGNFLVWET----
DQ246083        GTLKLVGGHYNFVSGEFLTWDNKQPS
Repeat_A        GTLKLVGAHYDFVKGQFVTWEPP---
                *******.**:**.*:*:.*:
```

Fig. 3.17. *ClustalW amino acid alignment of the AY109272 and DQ246083 translated sequences with the maize CA amino acid sequence of Repeat A.*

3.4 Discussion

3.4.1 The maize CA2 gene

Approximately 7.5 kb corresponding to the CA2 gene from maize had been sequenced and contained many small exons characteristic of plant genes (Burnell, 2000). It contained the first nine exons encoding CA2, eight of which encoded the first protein domain, Repeat A. The last exon in the sequence encoded the start of the second protein domain, Repeat C, which was identical to the first. Additionally 1,500 bp upstream of the first exon was sequenced.

3.4.1.1 Screening the maize genomic DNA library

To obtain the full-length CA2 gene, a maize genomic DNA library was screened with a probe corresponding to CA2 cDNA sequence (Genbank accession U08401, gi:606810). The genomic clone obtained was purified and further analysed to allow sequencing and assembly of a full-length gene fragment. However, the complexity of assembling highly repetitive sequences meant that only two shorter assemblies were produced that were 3.55 kb and 2.7 kb (Fig. 3.7). The 3.55 kb assembly contained sequence corresponding to the unique 276 bp insert of the CA2 cDNA, positively confirming identification of CA2 genomic sequence, and the 2.7 kb assembly provided sequence data for the 3′-end of the CA2 gene.

Alignment of the 7.5 kb CA2 gene sequence with the genomic DNA library clone showed many sequence variations (Appendix 3.3). These variations included nucleotide substitutions, deletions of up to 100 bp and insertions of up to 50 bp, predominantly in intron regions. However, the sequence of the exons was similar with exon-intron boundaries conserved (Section 4.3.1.1, Chapter 4). These variations, taken together with the high level of allelic polymorphism of the maize genome, suggest that it is likely that the clone obtained from the library screen was not the same allele as that already sequenced. Additionally, the source of the genomic DNA library was a Missouri 17 inbred line while the 7.5 kb sequence was generated from genomic DNA sourced from a Golden Bantam cultivar, and even between cultivars there can be a large degree of polymorphism (Bortiri *et al.,* 2006).

The most significant difference in coding sequence between the 7.5 kb sequence and the genomic DNA library clone was in the second exon where two insertions of two and four nucleotides would create a change in the open reading frame as this sequence was translated. This may suggest the gene isolated is actually a pseudogene although numerous amplifications of cDNA across this region support transcription of the CA2 gene including that part of sequence (Section 6.3.3, Chapter 6).

3.4.1.2 Southern analysis of genomic DNA

To determine the number of beta-CA genes in maize, a Southern blot was performed using genomic DNA that had been modified by two restriction enzymes, *Eco*RI and *Kpn*I (Fig. 3.8). As sequence data was available, the number and size of hybridizing fragments could be predicted for genomic DNA treated with *Eco*RI (Table 3.4). In contrast, to enable identification of total gene number *Kpn*I was used as there were no recognition sites for that enzyme within the CA2 gene sequence or the cDNA sequences of CA1 and CA3.

Table 3.4. Expected CA2 gene fragments after modification with *Eco*RI.

Size (kb)	Fragment description
2.36	Upstream region, Exon 1, 5′-end of Intron 1
1.72	3′-end of Intron 1, Exon 2 to Intron 3
2.83	Intron 3 to Intron 9
2.37+	Intron 9 to next *Eco*RI recognition site in genomic sequence
9.28	**Total**

Two different probes were used, the first contained coding sequence across several exons in the repeating portion of the sequence encoding Repeat A (*Repeat* probe), and the second corresponded to the second exon, which is the unique 276 bp insert of the CA2 sequence (*Insert* probe). The expected hybridizing fragment sizes for *Eco*RI treated genomic DNA could be determined for each probe. A 1.72 kb fragment would hybridize with the

Insert probe, while two bands would be expected to hybridize with the *Repeat* probe, 2.83 kb and another band larger than 2.37 kb (the next *Eco*RI recognition site in the genomic DNA sequence was unknown).

The results of the Southern using the CA2 genomic DNA library clone were also taken into consideration. In order to facilitate analysis of the genomic DNA library clone, it was treated with a number of different restriction enzymes, and the resulting hybridizing fragments were subsequently chosen for further sub-cloning and sequencing (Fig. 3.6). The probe used for this Southern analysis had sequence corresponding to both the insert (Exon 2) and to the Repeat A cDNA sequence, and the size of the hybridizing bands were determined and compared to those observed for the genomic DNA Southern (Table 3.5).

Table 3.5. Hybridizing DNA fragments observed from Southern analysis of the CA2 genomic DNA library clone and the genomic DNA Southern (Fig. 3.8).

		Fragment size (kb)	
		Genomic DNA Southern	CA2 library clone Southern
Restriction Enzyme	**Probe**		
*Eco*RI	*Repeat*	8.94	3.4
		5.34	2.4
		3.37	1.8
	Insert	7.89	
		5.18	
		4.12	
		1.91	
*Kpn*I	*Repeat*	20.58	10 +
		8.22	
		7.58	
	Insert	20.58	

It was expected that an *Eco*RI fragment of approximately 1.7 kb would hybridize with the *Insert* probe in the genomic DNA Southern, and the 1.9 kb fragment observed was most likely the corresponding fragment, also identified in the CA2 library clone Southern at 1.8 kb (Table 3.5). This fragment would contain Exon 2 of the CA2 gene sequence as well as intron sequence on either side of the exon. Additionally, an *Eco*RI fragment of approximately 2.83 kb should hybridize with the *Repeat* probe, and the closest hybridizing fragment observed in both the genomic DNA Southern and the CA2 library clone Southern was 3.4 kb. This fragment would contain the exons encoding Repeat A of the CA2 gene sequence, the introns of this sequence, and the first exon of Repeat C. The fragment size was larger than expected by approximately 600 bp, and the discrepancies in fragment lengths could be attributable to inaccuracies in molecular size predictions. Alternatively, the difference in molecular weight may be due to an insertion in this region of the gene that exists in the maize cultivar used to extract the genomic DNA, which may not exist in the cultivars used to generate the original 7.5 kb sequence or the genomic DNA library.

A second *Eco*RI fragment was predicted to hybridize with the *Repeat* probe on the genomic DNA Southern and was expected to be larger than 2.4 kb, with the subsequent *Eco*RI recognition site in the genomic DNA sequence not known. Two *Eco*RI fragments remained unaccounted for on the genomic DNA Southern at 8.94 kb and 5.34 kb, with one of these fragments therefore corresponding to the 3′-end of the CA2 gene (Fig. 3.8). A hybridizing fragment of 2.4 kb is observed in the CA2 genomic library clone Southern, and this fragment contained the sequence corresponding to the 3′-end of the CA2 gene.

The implication from the results obtained with the *Repeat* probe is that there are at least two copies of a gene encoding CA in the maize genome, one of which would be the CA2 gene. This was supported by the *Kpn*I fragments that hybridized with the *Repeat* probe. Two bands were observed that have molecular weights of approximately 7.2 kb and 8.6 kb. Furthermore, there was strong hybridization at approximately 20 kb, which may indicate the presence of additional genes. This signal was also observed when the *Kpn*I-treated genomic DNA was hybridized with the *Insert* probe (Fig. 3.8).

There are several more fragments than expected of *Eco*RI treated genomic DNA that hybridized with the *Insert* probe. Specifically there are three fragments, 7.89 kb, 5.18 kb and

4.12 kb that were not explained by the restriction pattern expected based on the CA2 gene sequence. At least one of these may be associated with the second *Kpn*I fragment that hybridized with the *Repeat* probe, representing a second copy of the CA2 gene. It may be possible that this is a second locus, highly similar to the first, with differences accounted for by allelic variation. Therefore it remains that there are at least two other copies of genes that also contain sequence homologous to the second exon of the CA2 gene.

The ploidy of maize should be taken into consideration when considering the number of genes that may exist in the genome that encode CA. Increased ploidy will increase the possibility that allelic variations of the gene exist. These allelic variations may account for the similarity and differences firstly in the three maize isozymes, CA1, CA2 and CA3 (Burnell, 2000), but additionally account for the presence of other as yet uncharacterized beta-CA genes. The identification of two CA genes in sorghum was also based on Southern analysis of genomic DNA using probes that corresponded specifically with two different transcript sequences that were isolated, as well as a probe corresponding to the conserved coding sequence (Wyrich *et al.,* 1998). The transcript-specific probes hybridized with fragments that appeared as component and complementary fragments in the pattern observed with the coding sequence probe. However, additional hybridizing bands were observed when the coding sequence probe was used potentially indicating the presence of more beta-CA gene copies.

3.4.2 Interrogation of the Maize Genomic Databases

The genomic DNA sequence was used to interrogate available databases (Table 3.1) identifying several assemblies with homology to the CA2 gene sequence. Specifically, two assemblies that appeared adjacently positioned, AZM4_68974 and AZM4_68973, provided the full length sequence of the CA2 gene (Fig. 3.12 and Fig. 3.14). The full-length gene spanned approximately 7 kb of genomic DNA and consisted of 14 exons interrupted by 13 introns. The first two exons were found in the AZM4_68974 assembly, including the 276 bp insert unique to the CA2 cDNA sequence (Burnell and Ludwig, 1997). The second assembly AZM4_68973, consisted of two nearly identical sets of six exons, corresponding to the CA2 gene. These exons encode two identical protein domains, which have been designated Repeat A and Repeat C respectively (Burnell and Ludwig, 1997).

Additionally, two other apparent beta-CA genes were identified, AZM4_38976 and AZM4_23203. Unlike the CA2 gene, these genes were only approximately 3.5 kb and encoded only a single protein domain. Neither of these nucleotide sequences contained an *Eco*RI or a *Kpn*I recognition site, although the sequences were homologous with the CA2 gene sequence and may be represented by the 20 kb hybridizing band observed by Southern blotting (Fig. 3.8). While the corresponding mRNA sequence for the AZM4_38976 assembly could not be identified, the AZM2_23203 assembly encoded for Genbank entry AY109272 (version gi:21212748, Appendix 3.7), which was 978 bp. This mRNA entry was one of several for CA in maize (NCBI Unigene: Zm.93944, Zm.93673 and Zm.78683), and was sourced from an expressed sequence tag collection (Gardiner *et al.,* 2004). The Unigene entry Zm.93944 included the two cDNA sequences identified as CA1 and CA2 (U08401 and U08403; Burnell and Ludwig, 1997), as well as three other partial mRNA sequences. A third Unigene entry, Zm.78683, included Genbank entry DQ246083, and like AY109272 appeared to be full length. The AY109272 and DQ246083 translated sequences were 91% homologous, and were also 86% and 83% homologous with Repeat A (Fig. 3.17).

Using the resources of the maize genomics database (www.maizegdb.org), three entries were found when a search was performed with the terms ‘carbonic anhydrase’. Like the NCBI database, no other full length coding sequences have been reported apart from CA1 and CA2 (Burnell and Ludwig, 1997). The entry indicated that the gene(s) encoding CA mapped to both chromosomes 3 and 8, which are homologous in the maize genome. This is supported by restriction fragment length polymorphism (RFLP) mapping using sorghum, which located CA to chromosomes 3 and 8, and by comparative mapping to chromosome 1 in rice (Wyrich *et al.,* 1998). Despite the identification of two different transcripts in sorghum, they could not be resolved on a genetic map as only one transcript showed length polymorphism.

3.4.3 The CA2 gene encodes two identical protein domains

The AZM4_68973 assembly consisted of 12 exons of the CA2 gene sequence, the first six of which were nearly identical to the second group of six. As well as having over 90% sequence similarity at the nucleotide and amino acid level, the exon-intron pattern observed is also similar, with two small exons being separated from the subsequent four exons by a large intron (Section 4.3.1.1, Chapter 4). The CA2 gene containing the exons in this arrangement may be a result of gene duplication. The generation of pseudo, partial and chimeric gene copies in the maize genome due to the action of retro-transposons is well documented (Song and Messing, 2003; Zhang *et al.,* 2006; Ilic *et al.,* 2003), and gene duplication would not be surprising. However, the length of CA transcripts in other monocot species such as sorghum and sugarcane is also indicative of the presence of more than one protein domain encoded by each transcript. As the CA2 gene in maize encodes two protein domains, the question arises of whether both domains are required for catalytic activity, or whether the mRNA is processed and translated into discrete isozymes (Section 5.3.2.1, Chapter 5).

The length of CA-encoding transcripts appears to be a characteristic unique to C_4 monocot species, although other dual-domain CAs have been identified. The unicellular red alga, *Porphyridium purpureum* has a 62 kDa beta-CA that appears to be a product of gene duplication, with the N- and C-terminal halves of the mature enzyme having sequence homology as well as two active sites (Mitsuhashi and Miyachi, 1996). This arrangement has also been observed in the alpha-CA gene family, with a 63 kDa enzyme from *Dunaliella salina* shown to have two active sites (Moroney *et al.,* 2001), and a dual-domain CA has been identified in the gigantic clam *Tridacna gigas.* However, in this case the C- and N-terminal CA domains have little amino acid homology at 29% (Leggat *et al.,* 2005).

3.5 Conclusion

The CA2 gene was isolated from a maize genomic DNA library using a probe containing the 276 bp insert unique to CA2. Two assemblies were created that corresponded to the 5′-end of the CA2 gene, including the 276 bp insert, as well as sequence corresponding to the 3′-end of the CA2 gene, which included the last three exons, the 221 bp 3′-untranslated region and several hundred nucleotides downstream. While the full length gene sequence was not determined, a different sequencing technique may have provided the means to overcoming the difficulties associated with assembling highly repetitive genomic DNA sequence. A method that has been used with success is transposon-facilitated sequencing. This process involves creating a library of artificial transposon insertions in the clone of interest, which provide unique primer annealing sites and allow mapping by PCR (Strathmann *et al.,* 1991; Devine *et al.,* 1997).

The question of the number of CA genes present in the maize genome follows from the finding of more than one CA transcript in maize, as well as several other plant species (Burnell and Ludwig, 1997; Wyrich *et al.,* 1998; Tetu et al., 2007). Whether these transcripts are encoded by more than one beta-CA gene in maize was determined by Southern blotting using CA cDNA probes, as well as using specific primer binding sites to amplify over introns. The results obtained confirmed that there are as many as three distinct CA genes present in maize. The next objective for this analysis would be to screen the maize genomic DNA library to obtain sequence data corresponding to the other CA genes present in maize, and correlate this data with the function of the genes identified.

While CA plays an important role in the cytosol of mesophyll cells providing the bicarbonate substrate for PEP carboxylase, the requirement for more than one CA isozyme in the C_4 photosynthetic pathway, and therefore more than one gene, has not been identified (Hatch and Burnell, 1990). Many photosynthetic genes that encode proteins involved in the C_4 photosynthetic pathway also have non-photosynthetic functions, or have evolved from C_3 plants where the function and location of these enzymes has changed (Ku *et al.,* 1996). The presence of more than one CA gene indicates that CA also plays a physiological role in maize.

Further analysis of the maize genome by genomic DNA library screening could also indicate whether the maize CA gene family is affected by the action of retro-transposons. The maize genome is complex, with a low proportion of gene-coding sequence (7%), and a high proportion of retro-transposon elements (Messing *et al.,* 2004; SanMiguel and Bennetzen, 1998). These retro-transposons are thought to generate insertions, deletions and pseudo-gene copies. Whether the maize CA gene family contains partial or chimeric gene copies could therefore be determined by isolating CA genes and analysing sequences to provide more information on the structural organization of the CA gene family.

During the course of this investigation, the maize genome sequencing projects have progressed with the sequences of several CA genes available in genomic databases (Table 3.1). One of these genes was composed of two assemblies (AZM4_68974 and AZM4_68973; www.maizegenome.org) and corresponded to the CA2 gene. This assembly contained the 276 bp insert (Exon 2), which is unique to the CA2 cDNA sequence (Genbank accession: U08401, version: gi:606810), as well as two nearly identical domains, encoding Repeat A and Repeat C. Whether both of these domains are required for enzymatic activity of the translated protein was investigated further (Chapter 5), considering that two assemblies were also found, AZM4_38976 and AZM5_23203 (www.maizegenome.org) that appeared to contain only one domain.

CHAPTER 4

4. Analysis of the Maize CA Nucleotide Sequence

Introduction

CA in maize is encoded by a multi-gene family, resulting in the expression of several isozymes within the plant. Analyzing the primary structure of maize CA provided a means to obtain information related to the expression, function and localization of these isozymes. This is of particular interest as the primary structure of these isozymes has several unique features that do not appear in CAs from other plant species, such as pea, spinach and rice. With the availability of both genomic DNA and cDNA sequences, an investigation into gene expression regulation was possible based on identification of putative regulatory elements and gene structure analysis.

4.1.1 CA in C_4 monocot plant species

The primary structures of CA from a sub group of plants including maize, sugarcane and sorghum have unique characteristics. These plants are grasses and are classified as monocots, developing with only one seed leaf (Fig. 4.1). Species such as rice and *Hordeum vulgare* (barley) are also included in this group however maize, sugarcane and sorghum use the C_4 photosynthetic pathway resulting in increased efficiency of carbon fixation and decreased photorespiration (Hatch *et al.,* 1967).

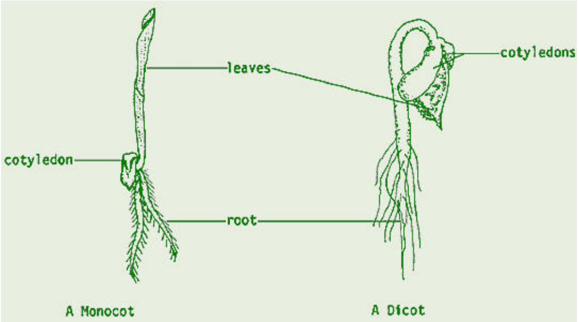


Fig. 4.1. *Structural differences between monocot and dicot plants (http://biology.clc.uc.edu/graphics/bio106/bean.jpg).*

The CA isozymes of certain C_4 monocot species are encoded by comparatively large transcripts, due to the presence of repeating sequences (Burnell and Ludwig, 1997; Wyrich *et al.,* 1998). This characteristic appeared to be specific for NADP-malic enzyme (NADP-ME) type C_4 monocot species, as the CA isozymes from other C_4 monocot species such as *Urochloa panicoides* were not composed of repeating sequences, but consisted of only one protein-encoding domain (Burnell, 2000).

4.1.2 C_4 photosynthesis and NADP-ME catalyzed decarboxylation in bundle sheath cells

Plants that use the C_4 photosynthetic pathway have two cell types separately involved in the processes of carbon assimilation and carbon fixation, which are the mesophyll cells and bundle sheath cells surrounding the vascular tissue (Edwards *et al.,* 1985). In the mesophyll cells carbon dioxide from the atmosphere is converted to bicarbonate by CA present in the cytosol. This bicarbonate then serves as the substrate for PEP carboxylase in the first reaction of the C_4 pathway, which is the carboxylation of PEP to produce four carbon acids, such as oxaloacetate and malate. These four carbon acids diffuse into the bundle sheath cells, where the enzymes of the Calvin-Benson cycle are located (Weiner *et al.,* 1988).

In the bundle sheath cells the four carbon acids are decarboxylated releasing carbon dioxide that is used by Rubisco, a large multi-subunit enzyme that fixes carbon dioxide into a three carbon compound through ribulose-1,5-bisphosphate carboxylation (Lorimer, 1981).

Rubisco is able to use both carbon dioxide and oxygen as substrates to function as either a carboxylase or an oxygenase, and the efficiency of the C_4 pathway is a consequence of the effective concentration of carbon dioxide around Rubisco, precluding the energetically wasteful oxygenase reaction. There are three different types of C_4 plants, classified according to the enzyme used to decarboxylate the four-carbon compound in the bundle sheath cells (Edwards *et al.,* 1985). These enzymes include NADP-ME, NAD-malic enzyme (NAD-ME), and PEP carboxykinase (PEP-CK), each of which has a different sub-cellular location. NADP-ME is located in the chloroplast, NAD-ME in the mitochondria and PEP-CK is located in the cytosol of the bundle sheath cells (Furumoto *et al.,* 1999).

While sub-classification of C_4 plants is based on the primary decarboxylating enzyme that is used to supply carbon dioxide to Rubisco, all three enzymes are found in C_4 plants. PEP-CK is involved in gluconeogenesis and providing C_4 acids through the reverse carboxylation reaction in a range of organisms (Furumoto *et al.,* 1999). In NADP-ME type C_4 plants, PEP-CK in the cytosol is involved in amino acid metabolism, for example the decarboxylation of asparagine (Lea *et al.,* 2001; Walker *et al.,* 2002). In the PEP-CK type C_4 plant, *U. panicoides,* NAD-ME in the mitochondria of bundle sheath cells is also involved in decarboxylation of four-carbon compounds, in particular malate, regenerating PEP for the initial carboxylation reaction that occurs in the mesophyll cells (Burnell and Hatch, 1988). NADP-ME is present in both plants and animals where it is involved in a wide range of cellular processes from generating reducing power for fatty acid synthesis to regulating intracellular pH (Rothermel and Nelson, 1989). In maize, the inorganic carbon substrate for Rubisco is produced by NADP-ME, in a reaction also producing pyruvate from the oxidation of malate coupled with the reduction of $NADP^+$ (Hausler *et al.,* 1987). The enzyme is located in the chloroplasts of the bundle sheath cells, where the pH favours oxidative decarboxylation.

4.1.3 Gene structure of CA isozymes from NADP-ME type C_4 monocot species

The CA transcripts in maize, sugarcane and sorghum are much longer than those from other C_4 monocots, such as the CA transcripts from *U. panicoides* (Burnell, unpublished; Genbank accessions U19739 and U19741), and CA transcripts from C_3 species such as rice where only one CA isozyme is present (Suzuki and Burnell, 1995). Longer transcripts were identified in the C_4 monocot sorghum (Wyrich *et al.,* 1998), and in sugarcane (Burnell, unpublished), but not in the C_4 dicot *Flaveria bidentis,* where there are three transcripts that have open reading frames of less than 1,000 bp (Tetu *et al.,* 2007). In rice, a C_3 monocot species, the CA transcript has an open reading frame of 819 bp (Suzuki and Burnell, 1995).

In maize, there are three CA transcripts that are 1.3 kb, 1.9 kb and 2.2 kb, while in sorghum the CA transcripts are 1.2 kb, 1.7 kb and 2.1 kb (Burnell and Ludwig, 1997; Wyrich *et al.,* 1998). The corresponding cDNAs in maize have 600 bp repeating sequences that encode nearly identical protein domains, which have been designated Repeat A, Repeat B and Repeat C (Fig. 4.2). The longest CA transcript identified in maize, CA1, had three repeating sequences, Repeat A, Repeat B and Repeat C, while CA2 had only Repeat A and Repeat C (Burnell and Ludwig, 1997; Burnell, 2000).

```
Repeat_B        CCCCAGGACGCCATCGAGCGCTTGACGAGCGGCTTCCAGCAGTTCAAGGTCAATGTCTAT
Repeat_C        CCCCAGGACGCCATCGAGCGCTTGACGAGCGGGTTCCAGCAGTTCAAGGTCAATGTCTAT
Repeat_A        ATGGACCCCACCGTCGAGCGCTTGAAGAGCGGGTTCCAGAAGTTCAAGACCGAGGTCTAT
                    *   * ** ************ ****** ****** ********  * * ******

Repeat_B        GACAAGAAGCCGGAGCTTTTCGGGCCTCTCAAGTCCGGCCAGGCCCCCAAGTACATGGTG
Repeat_C        GACAAGAAGCCGGAGCTTTTCGGGCCTCTCAAGTCCGGCCAGGCCCCCAAGTACATGGTG
Repeat_A        GACAAGAAGCCGGAGCTGTTCGAGCCTCTCAAGTCCGGCCAGAGCCCCAGGTACATGGTG
                ***************** **** *******************  ***** **********

Repeat_B        TTCGCCTGCTCCGACTCCCGTGTGTGCCCGTCGGTGACCCTGGGCCTGCAGCCCGGCGAG
Repeat_C        TTCGCCTGCTCCGACTCCCGTGTGTGCCCGTCGGTGACCCTGGGCCTGCAGCCCGGCGAG
Repeat_A        TTCGCCTGCTCCGACTCCCGCGTGTGCCCGTCGGTGACACTGGGACTGCAGCCCGGCGAG
                ******************** ***************** ***** ***************

Repeat_B        GCCTTCACCGTTCGCAACATCGCCGCCATGGTCCCAGGCTACGACAAGACCAAGTACACC
Repeat_C        GCCTTCACCGTTCGCAACATCGCCGCCATGGTCCCAGGCTACGACAAGACCAAGTACACC
Repeat_A        GCATTCACCGTCCGCAACATCGCTTCCATGGTCCCACCCTACGACAAGATCAAGTACGCC
                ** ******** ***********  ***********  *********** ******* **

Repeat_B        GGCATCGGGTCCGCCATCGAGTACGCTGTGTGCGCCCTCAAGGTGGAGGTCCTCGTGGTC
Repeat_C        GGCATCGGGTCCGCCATCGAGTACGCTGTGTGCGCCCTCAAGGTGGAGGTCCTCGTGGTC
Repeat_A        GGCACAGGGTCCGCCATCGAGTACGCCGTGTGCGCGCTCAAGGTGCAGGTCATCGTGGTC
                ****  ******************** ******** ********* ***** ********

Repeat_B        ATTGGCCATAGCTGCTGCGGTGGCATCAGGGCGCTCCTCTCCCTCAAGGACGGCGCGCCC
Repeat_C        ATTGGCCATAGCTGCTGCGGTGGCATCAGGGCGCTCCTCTCACTCCAGGACGGCGCACCT
Repeat_A        ATTGGCCACAGCTGCTGCGGTGGCATCAGGGCGCTCCTCTCCCTCAAGGACGGCGCGCCC
                ******** ******************************** *** ********** ** 

Repeat_B        GACAACTTCCACTTCGTGGAGGACTGGGTCAGGATCGGCAGCCCTGCCAAGAACAAGGTG
Repeat_C        GACACCTTCCACTTCGTCGAGGACTGGGTTAAGATCGGCTTCATTGCCAAGATGAAGGTA
Repeat_A        GACAACTTCCACTTCGTGGAGGACTGGGTCAGGATCGGCAGCCCTGCCAAGAACAAGGTG
                **** ************ *********** * *******  *  ********  ***** 
```

```
Repeat_B        AAGAAAGAGCACGCGTCCGTGCCGTTCGATGACCAGTGCTCCATCCTGGAGAAGGAGGCC
Repeat_C        AAGAAAGAGCACGCCTCGGTGCCGTTCGATGACCAGTGCTCCATTCTCGAGAAGGAGGCC
Repeat_A        AAGAAAGAGCACGCGTCGGTGCCGTTCGATGACCAGTGCTCCATCCTGGAGAAGGAGGCC
                ************** ** ************************** ** ************

Repeat_B        GTGAACGTGTCGCTCCAGAACCTCAAGAGCTACCCCTTCGTCAAGGAAGGGCTGGCCGGC
Repeat_C        GTGAACGTGTCCCTGGAGAACCTCAAGACCTACCCCTTCGTCAAGGAAGGGCTTGCAAAT
Repeat_A        GTGAACGTGTCGCTCCAGAACCTCAAGAGCTACCCCTTCGTCAAGGAAGGGCTGGCCGGC
                *********** **  ************ ************************ **

Repeat_B        GGGACGCTCAAGCTGGTTGGCGCCCACTACGACTTCGTCAAAGGGCAGTTCGTCACATGG
Repeat_C        GGGACCCTCAAGCTGATCGGCGCCCACTACGACTTTGTCTCAGGAGAGTTCCTCACATGG
Repeat_A        GGGACGCTCAAGCTGGTTGGCGCCCACTACGACTTCGTCAAAGGGCAGTTCGTCACATGG
                ***** ********* * ***************** ***  ***  ***** ********

Repeat_B        GAGCCT
Repeat_C        AAA---
Repeat_A        AGCCT-
```

Fig 4.2. *ClustalW nucleotide alignment of Repeat A, Repeat B and Repeat C of the maize CA cDNA sequences.*

Repeats A and C in CA1 and CA2 are identical and are homologous with Repeat B (Fig. 4.2). Furthermore, Repeat B may be a chimeric product composed of Repeats A and C, with the 5′-end of Repeat B having homology with the 5′-end of Repeat C, while the 3′-end of Repeat B is more similar to the nucleotide sequence at the 3′-end of Repeat A.

The CA cDNAs also contain a 150 bp 5′-leader sequence that is distinct from the repeating sequences. CA2 has a unique 276 bp insert after the 5′-leader sequence but before the repeating sequences that is not present in CA1 or the shortest CA transcript (1.3 kb). Whether the leader sequence or the 276 bp insert are translated and form part of the active protein or are involved in post-transcriptional regulation of expression remains to be determined. Analysis of genomic sequence data reveals this insert is encoded by one exon, while the exons encoding the repeating sequences are considerably smaller (Section 3.3.1.1, Chapter 3).

The third CA cDNA corresponding to the smallest CA transcript has been characterized (Burnell, unpublished). As this is the smallest transcript, it has been designated CA1, and also contains the 5′-leader sequence in addition to Repeats A and C. The longest transcript (containing Repeats A, B and C) has since been designated CA3 while CA2, containing the unique 276 bp insert, remains CA2 (Burnell, 2000).

4.1.4 Rationale

This purpose of this study was to analyze the nucleotide sequence data of the beta-CA gene family from maize, an NADP-ME type C_4 monocot in order to identify the molecular structure and location of the maize CA isozymes. The genomic sequence data was also interrogated to determine how gene expression is regulated. This included responsiveness to environmental factors as well as regulation of mRNA stability and translation. Regulation may be at the level of transcription or be due to post-translational modifications, as the size of the CA subunits predicted from the deduced amino acid sequences do not correspond with immuno-reactive bands when maize leaf extract is hybridized with CA antibody (Burnell and Ludwig, 1997; Section 6.3.4.1, Chapter 6). The transcripts or the translated protein may be subject to as yet unidentified modifications. The location and expression levels of the CA transcripts were analyzed by amplification and quantification using different maize tissues allowing the relative abundance of the CA transcripts to be examined.

Materials and methods

4.2.1 Generation of a phylogenetic tree

CA sequences (Roeske and Ogren, 1990; Majeau and Coleman, 1992; Suzuki and Burnell, 1995; Burnell *et al.*, 1990a; Burnell and Ludwig, 1997) were assembled and aligned using MacVector^{TM} 7.0 and ClustalW sequence analysis software (Thompson *et al.,* 1994).

4.2.2 Analysis of the CA2 gene structure

4.2.2.1 Identification of intron/exon boundaries

Intron/exon boundaries were located by comparison of the cDNA and genomic DNA sequences. Sequences were assembled and aligned using MacVector^{TM} 7.0 and ClustalW sequence analysis software (Thompson *et al.,* 1994).

4.2.2.2 Analysis of nucleotide composition of the CA gene in maize and rice

To calculate the percentage of G/C nucleotides composing the exons and introns of the CA2 gene, the total number of nucleotides was counted and the number of G and C nucleotides was expressed as a percentage of the total number.

The rice CA genomic DNA sequence data was available from The Institute for Genomic Research database (www.tigr.org). The amino acid sequence of the rice CA enzyme was obtained from the NCBI database (http://www.ncbi.nlm.nih.gov/; accession number AAA86943, version gi:606817).

4.2.2.3 Identification of CpG islands

Nucleotide sequence up to 1,500 bp upstream of the first exon for both the maize (AZM4_68974) and rice CA gene sequences were analyzed to identify the presence of putative CpG islands. The CpG island predictor database at http://www.ebi.ac.uk/emboss/cpgplot/ was used for this analysis.

4.2.3 Transcription initiation site

4.2.3.1 Determining the transcription initiation site by 5′-rapid amplification of cDNA ends (5′-RACE)

Total RNA was isolated from maize leaves (Section 2.1.4, Chapter 2), and used as template for the first strand synthesis reaction (SMART^{TM} RACE cDNA Amplification Kit, Clontech), generating 5′-RACE-ready cDNA. The cDNA was diluted ten-fold with Tricine-EDTA buffer (10 mM Tricine-KOH pH 8.5, 1 mM EDTA) and 2.5 µl used as the template for 5′-RACE PCR using four gene-specific primers (Table 4.1) in a 50 µl reaction.

Table 4.1. Sequence of the gene-specific primers used for 5′-RACE PCR.

Primer	**Sequence**
CA7	5′-ttcaagcgctcgacggt-3′
CA1R	5′-acgatgctggatgtggtg-3′
InsR	5′-gcccttggaggaagccttggaggg-3′
CA3	5′-tgagggtcccatttgcaagcc-3′

Amplification was performed using Advantage® 2 DNA polymerase (1.0 unit) and 1x PCR buffer (Clontech), with 0.2 mM dNTPs, 0.5 µM of each primer, and 1x universal primer mix (Clontech). Products generated by the 5′-RACE reaction (Table 4.2) were analyzed by electrophoresis on a 1% (w/v) agarose gel.

Table 4.2. Reaction cycling conditions used for 5′-RACE PCR.

Cycling Conditions:			
	Step 1	94ºC	30 s
	Step 2	65ºC	30 s
10 cycles from Step 1	Step 3	72ºC	2 min
	Step 4	94ºC	30 s
	Step 5	68ºC	30 s
25 cycles from Step 4	Step 6	72ºC	2 min

To increase the amount of amplified product obtained after the initial 5′-RACE reaction, a secondary reaction was performed using 3 µl of the original reaction as template. The products of the reaction were visualized by agarose gel electrophoresis and excised from the gel. The DNA fragments were purified using the QIAGEN gel extraction kit and protocol and used directly in a ligation reaction (Section 2.1.11, Chapter 2). Once sub-cloned, the DNA sequence of the PCR products was obtained and analysed (Section 2.1.13, Chapter 2).

4.2.3.2 PCR amplification of the 5′-end of the CA cDNA sequence

A maize cDNA library (Chastain Laboratory, Morehead University, USA) was used as template in a standard PCR (Section 2.1.14, Chapter 2). The reactions were performed using a gene-specific primer coupled with the T3 primer, which is located at the 5′-end of the library vector (pBluescript II SK(+/-), Stratagene). The sequence and position of the gene-specific primers, *CA1R, CA7* and *InsR* are shown in Table 4.1 and Fig. 4.3, and the cycling conditions in Table 4.3.

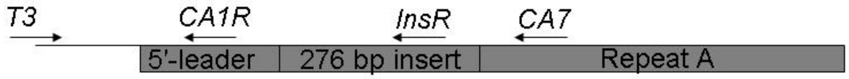


Fig. 4.3. *Schematic representation of the primer binding sites on the CA cDNA sequence in the three reactions used to identify the transcription initiation site using a maize cDNA library as template. Figure not to scale.*

Table 4.3. Cycling conditions used to amplify the 5′-end of the CA cDNA sequence from the maize cDNA library.

Cycling Conditions:			
	Step 1	95ºC	5 min
	Step 2	95ºC	45 s
	Step 3	48ºC	1 min
30 cycles from Step 2	Step 4	72ºC	1.5 min
	Step 5	72ºC	10 min

The products of the reactions were analysed by agarose gel electrophoresis, excised and extracted using a QIAGEN gel extraction kit and protocol and sub-cloned for sequence analysis.

4.2.4 Analysis of transcription factor binding sites

Putative transcription factor binding sites were identified in the region upstream of the initiating codon, as well as the first intron of the CA2 gene, using a plant promoter database (PlantProm DB at www.softberry.com.au) in conjunction with MacVector™ (Shahmuradov *et al.,* 2003).

Comparison of the maize CA2 gene sequence with the rice CA genomic sequence in the region up to 1,000 bp upstream of the initiating codon was performed using MacVector^{TM} to identify repetitive or homologous elements.

4.2.5 Semi-quantitative reverse transcriptase PCR

CA transcripts were amplified from cDNA made from RNA purified from along the length of the maize leaf (base, middle and tip) and from maize root tissue. Total RNA was purified (Section 2.1.4, Chapter 2), and cDNA generated using the QIAGEN QuantiTect Reverse Transcription Kit and protocol. Five separate isolations of RNA were performed, from which cDNA was generated to serve as the template for each reverse-transcriptase PCR (RT-PCR).

Reactions were performed with six primer pairs, including amplification of 18S rRNA, which served as a control (Table 4.4; Massonneau *et al.,* 2004). Three CA-specific primer combinations were used to enable isolation of specific regions of the CA transcript (Fig. 4.4). Also, a reaction was performed using a forward primer unique to one of the CA transcript sequences identified on the NCBI database, DQ246083 (http://www.ncbi.nlm.nih.gov), and a reaction was performed amplifying a region of the gene encoding pyruvate orthophosphate dikinase (PPDK).

Table 4.4. Sequence of the gene-specific primers used for semi-quantitative RT-PCR.

Reaction	Forward Primer	Reverse Primer	Product Size (bp)
18s rRNA	5′-atccctccgtagttagcttc-3′	5′-tgtcggccaaggctatatac-3′	108
Leader	5′-atgtacacattgcccgtccgtg-3′	5′-ttccggatgagcctgagcctg-3′	113
Repeat	5′-gtccatggtgttcgcctgctcc-3′	5′-cctagggaggctcccatgtgacg-3′	477
Insert	5′-ggcgggcataagagggg-3′	5′-gcccttggaggaagccttggaggg-3′	186
DQ246083	5′-cgaaggcgtacaattcatcc-3′	5′-ttctccaggatggagcactgg-3′	545
PPDK	5′- ttctggcaccggcgtgc-3′	5′- tgaggttcttcatggc-3′	144

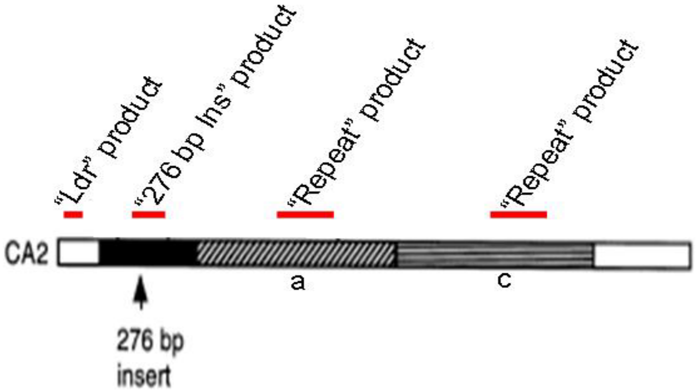


Fig. 4.4. Schematic representation of the regions of the CA transcript amplified. Ldr: amplification of the 5'-leader sequence, 276 bp Ins: amplification of a region of the unique 276 bp insert of CA2, Repeat: amplification within the repeating region of the CA transcript. Due to the repetitive sequence of the repeating regions, the primers used to amplify the repeat product would not distinguish between Repeat A, Repeat B or Repeat C. The CA2 cDNA schematic was taken from Burnell and Ludwig, (1997).

Five identical reactions from the same master mix of reagents were prepared. The master mix included 0.2 mM dNTPs (22 µl), 2.5 mM $MgCl_2$ (82.5 µl), 5x Promega PCR reaction buffer (220 µl) and 577.5 µl of double distilled water. To this 66 µl of each primer (5 µM) and 11 µl of Taq polymerase (Promega GoTaq Flexi, 5 units.µl^{-1}) were added, and

105 µl was aliquoted between ten 0.6 ml microcentrifuge tubes. 5 µl of each of the cDNA templates (at a dilution of 150 µg.ml^{-1}) was added to the master mix, 20 µl of which was aliquoted into individual reaction tubes. These were cycled simultaneously, and the reactions were stopped by removing from the thermo-cycler and placing on ice after 21, 24, 27, 30 and 33 cycles (Table 4.5).

Table 4.5. Cycling conditions used for amplification by RT-PCR.

Cycling Conditions:			
	Step 1	95ºC	2 min
	Step 2	95ºC	45 s
	Step 3	52ºC	45 s
33 cycles from Step 2	Step 4	72ºC	1 min

The PCR products generated were visualized with ethidium bromide staining after 1% (w/v) agarose gel electrophoresis. The relative abundance of each transcript was determined based on which cycle PCR product was first detected for each reaction. The relative abundance of transcript was confirmed after five repetitions (Table 4.6).

Table 4.6. Relative abundance of each transcript based on the PCR cycle at which the reaction product was first visible.

Cycle	Relative abundance score
21	10
24	8
27	6
30	4
33	2

Results

The morphological distinction of flowering plants into monocot and dicot types is supported by analysis of beta-CA gene sequence similarity (Fig. 4.5). There is greater CA sequence homology within monocot and dicot classes than between them, although the sequence homology between monocots and dicots is still significant (Burnell and Ludwig, 1997).

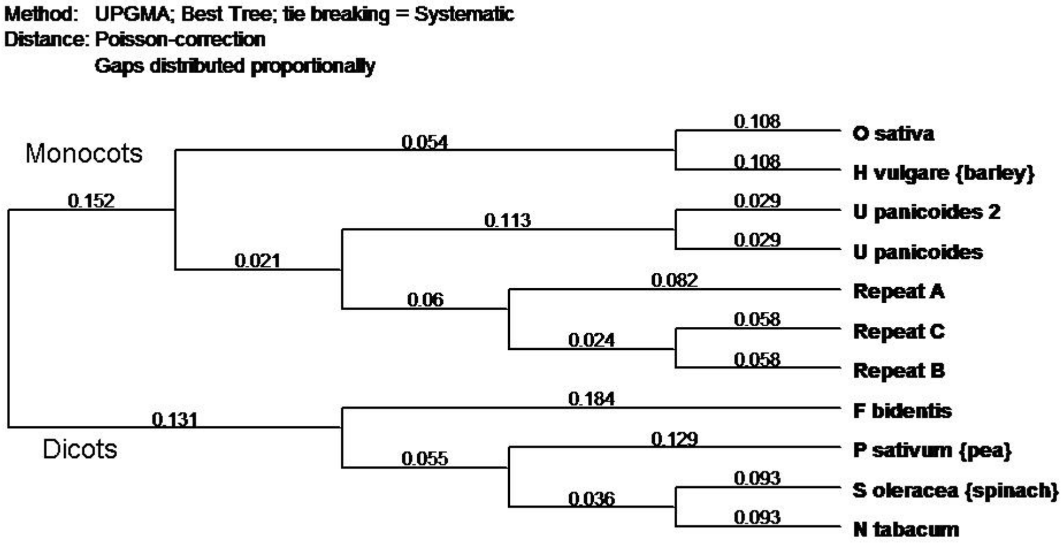


Fig. 4.5. *Phylogeny of the beta-CAs showing separation of monocot and dicot species. Repeat A, Repeat B and Repeat C refer to the maize CA amino acid sequences.*

4.3.1 The CA2 gene of maize

4.3.1.1 Gene structure

The exon-intron boundaries of the CA2 gene and one of the assemblies identified in the maize genome database (AZM4_23202; www.maizegenome.org, Section 3.3.2.3, Chapter 3) were identified by aligning cDNA and genomic DNA sequences (Table 4.7a and b). Predominantly, canonical sequences (*gt-ag)* are present at the exon-intron borders (Hanley and Schuler, 1988). The exception to this was the boundaries surrounding the second exon of the CA2 gene. The consensus sequence is present (*gtac*) at the end of the first exon, however the coding sequence obtained from analysis of the CA2 cDNA sequence indicated that this would place a stop codon (*taa)* at the start of the second exon.

Table 4.7a. Intron-Exon border sequences for the CA2 gene from maize.

	-EXON	intron-	Size (bp)	-intron	EXON-	
1	-CCGTCGT	gtacgtac-	1660	-catgttgct	AGTAAA-	2
2	-AGGGC	cgtcccctc-	380	-tgcagggc	TACATGG-	3
3	-GTCTATGA	gtaagtca-	130	-gttcgcag	CAAGAAG-	4
4	-GCCCCAG	gtacgcgc-	830	-cgctgcag	GTACATG-	5
5	-CGACAAG	gtacgtac-	97	-aactgcag	ATCAAG-	6
6	-ACAACTT	gtaagcag-	144	-atatccag	CCACTTG-	7
7	-GGAGAAG	gtacgtaa-	131	-ttctgcag	GAGGCCGT-	8
8	-GGAGCCT	gtaggggt-	863	-tcctgcag	CCCCAGGA-	9
9	-GTCTATGA	gtaagtca-	175	-tttcacag	CAAGAAG-	10
10	-CCCCAAG	gtatgcgc-	625	-gcatgcag	TACATGGT-	11
11	-CGACAAG	gtatatat-	94	-gactgcag	ACCAAGTA-	12
12	-CAACTT	gtaagtcg-	134	-gtacgcag	CCACTTC-	13
13	-GAGAAG	gtatgttg-	84	-ttctgcag	GAGGCC-	3′

Table 4.7b. Intron-Exon border sequences for the AZM4_23202 assembly (transcript AY109272, gi:21212748).

	-EXON	intron-	Size (bp)	-intron	EXON-	
1	-CAAGAAG	gtccgttc-	173	-cattgcag	GCCGCCAT-	2
2	-TGTATGA	gtaagtgc-	156	-tctcgcag	CAAGAAG-	3
3	-GCCCCCAAG	gtacgtga-	1615	-ccgtgcag	TACATGGT-	4
4	-TACGACAAG	gtaacgcg-	130	-gcatgcag	ACCAAGTA-	5
5	-GACAACTT	gtaagtat-	113	-gcatcgag	CCACTTCG-	6
6	-GGAGAAG	gtacgtac-	102	-tgctgcag	GAGGCCGT-	3′

This is not the case if the CA2 gene sequence codes for the other CA transcripts, CA1 or CA3 (Burnell and Ludwig, 1997; Genbank accession gi:606814). The coding sequence of CA1 and CA3 do not contain the second exon, with the first exon being followed directly by the third exon, corresponding to the start of Repeat A. In this case, the sequence (*gtg)* at the start of the first intron is present in the mRNA transcript coding for glycine at this position.

The gene structure of the maize and rice CA genes was also compared. There are general characteristics that are conserved between the maize and rice CA genes, such as the length of the intron between the first and second exons, and the pattern of two small exons being separated from the adjacent four exons by a relatively large intron. However, the maize CA gene contains two series of six exons, which encode repeating protein domains, while the rice CA gene only contains one (Fig 4.6).

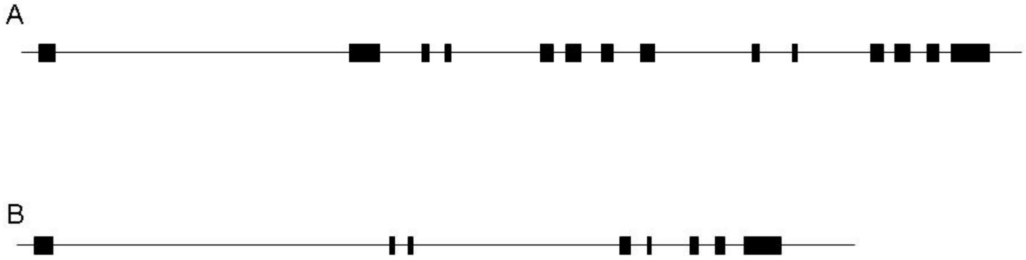


Fig. 4.6. *Schematic representation of (A) the maize CA2 gene and (B) the rice CA gene. Introns and exons are drawn to scale and indicated as lines or boxes respectively.*

4.3.1.2 Nucleotide composition

The maize CA2 gene was analysed for G/C composition and compared with the CA gene of rice (Fig. 4.7). Both genes were found to have a similar pattern of intron/exon length, as well as nucleotide composition of introns and exons. The exon regions were found to contain a higher percentage of G/C residues than the intron regions. Additionally, the percentage of G/C residues was higher at the 5′-end of the genes analyzed, and decreased towards the 3′-end.

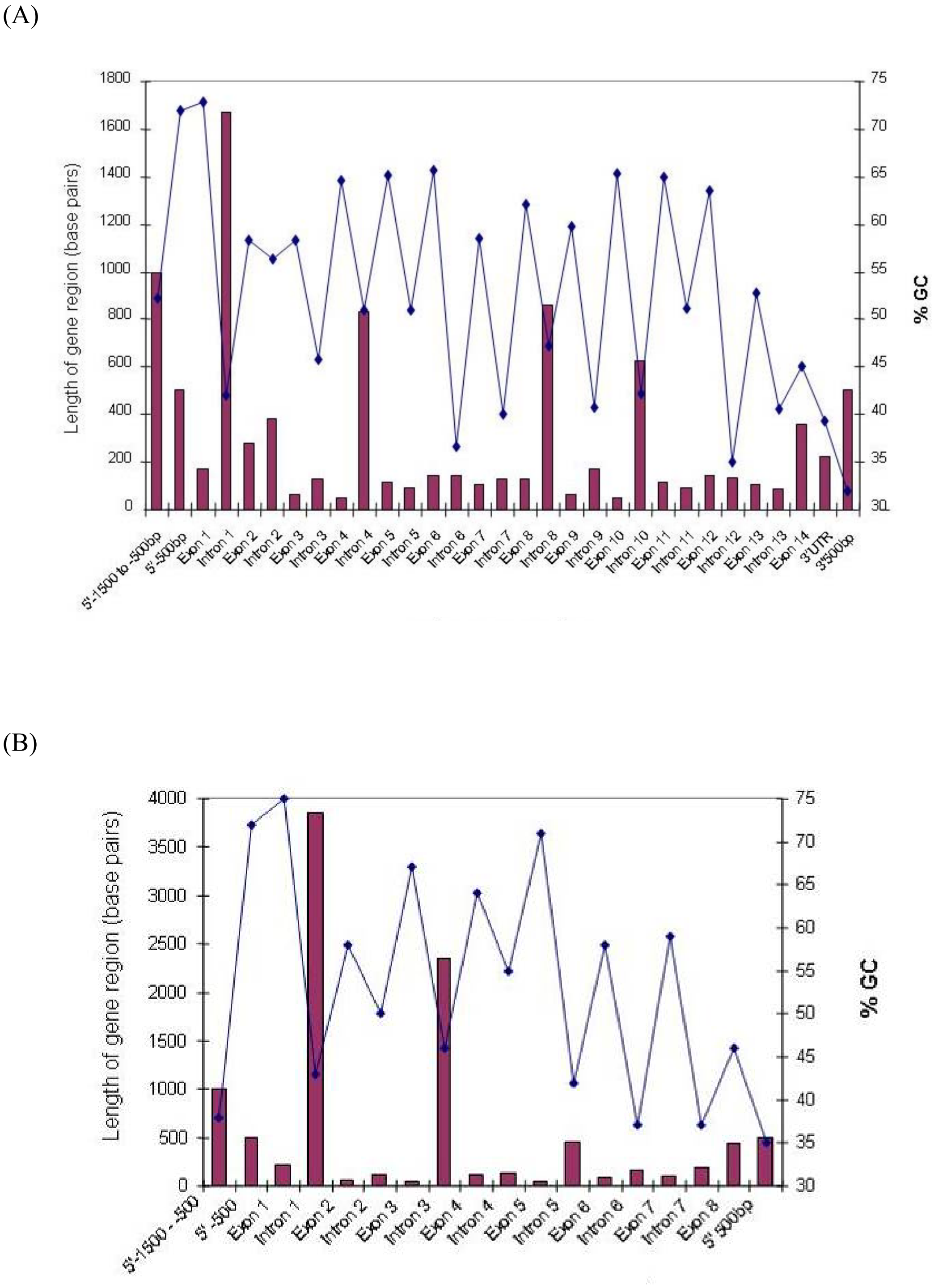


Fig. 4.7. *Analysis of gene structure for (A) the maize CA2 gene, and (B) the rice CA gene. The length of the introns and exons is compared, represented by the histograms, while the line graph shows the percentage of G/C residues composing these regions.*

4.3.1.3 Identification of CpG islands

The higher percentage of G/C residues at the 5′-end of the gene sequence was indicative of the presence of CpG islands in both the maize and rice genes. These were identified using a CpG island predictor database, which identified a 634 bp CpG island upstream of the initiating codon in maize (from -690 to -56 bp) and a 748 bp CpG island in rice (from -803 to -55 bp; Fig. 4.8).

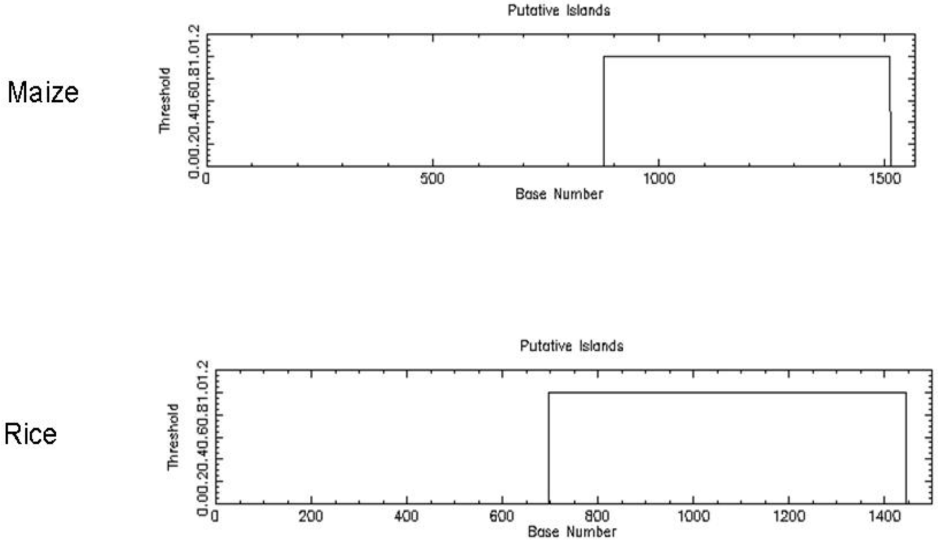


Fig. 4.8. *Location of CpG islands in the 5′-upstream region of the maize and rice CA gene sequences.*

4.3.2 Transcription initiation

4.3.2.1 Identification of the consensus sequence

The plant consensus transcription initiation sequence was identified in the CA2 gene sequence -379 to -390 bp upstream from the translation initiation site (Table 4.8; Matos *et al.,* 2001; Matsuoka, 1990). Initiation of transcription at this location would result in a 5′-untranslated region (UTR) of 394 bp.

Table 4.8. The transcription initiation site (indicated with *) consensus and the sequence around the cap site.

Gene	Sequence
Consensus	CTC*ATCA
PPDK	TTC*ACCA
PEP carboxylase	TTG*ATCA
CA	TTC*ATCA

4.3.2.2 Transcription initiation site

The 5′-RACE reaction was unsuccessful for generating the 5′-end of the maize CA cDNA sequence (results not shown). Additionally, PCR was performed using a maize cDNA library (Chastain Laboratory, Morehead University, USA) as template with three gene-specific reverse primers (Table 4.1), which were used separately in conjunction with the T3 primer, which binds the cDNA library vector (pBluescript II SK(+/-), Stratagene; Fig. 4.3). The products that were obtained using the primer combination described were analyzed by electrophoresis (Fig 4.9).

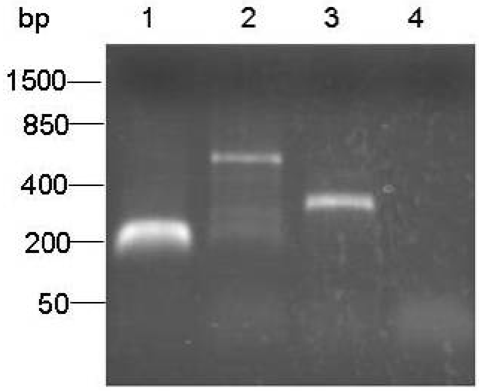


***Fig. 4.9.** Agarose (1.5%) gel electrophoresis of the products obtained from PCR on a maize cDNA library. Lane 1: CA1R and T3 primers; 2: InsR and T3 primers; 3: CA7 and T3 primers; and 4: negative control (minus template).*

The products of the three reactions were subsequently sub-cloned and the sequence analyzed. The sequences obtained for the products generated using the *CA1R*, *CA7* and two products of the *InsR* primers are shown in Table 4.9. In each instance the adjacent sequence corresponded to the linker/adaptor sequence used for generating the cDNA library (*cggcacgagg*; Stratagene).

Table 4.9. Alignment of the DNA sequences of PCR products generated to determine the transcription initiation site of the CA gene, compared with the genomic sequence at this position.

Source	Sequence
AZM4_68974	GATAAGCGGCACTCGCACGATCAATGTAC
CA1R product	GATCAATGTAC
InsR product 1	CTCGCACGATCAATGTAC
InsR product 2	CACGATCAATGTAC
CA7 product	CGCACGATCAATGTAC

4.3.3 Analysis of transcription regulation

The sequence upstream of the translation initiation codon, as well as the first intron of the CA2 gene were analysed for regulatory elements using the Regsite database (http://softberry.com), which is a plant-oriented collection of transcription regulatory elements, as well as the MacVector nucleotide subsequence analysis tool (Table 4.10, Appendix 4.1 and 4.2).

Table 4.10. Potential transcription factor binding elements found in the upstream region of the CA2 gene (5′-region, Appendix 4.1) and the first intron (Intron 1, Appendix 4.2). Sequence location is indicated relative to the AZM4_68974 sequence (5′-region) or the CA2 gene (Intron 1). The nucleotide position of each element is specified, with (-) indicating that the sequence appears on the reverse DNA strand.

	Element	Sequence	Location – 5′-region	Location – Intron 1	Function
1	Anaerobic responsive element (ARE) or GC-regulatory element	CCACG CCCCGG GCCGC	1525 (-) 1633 2433 1257 1740 (-) 2133 2401 1751 1998 2100 2138 2217 2220 2351 2361 2380	1165 1615 (-) 992 1543	Activation of maize alcohol dehydrogenase 1 gene in response to hypoxic stress. Proximal to transcription start (-99), may direct root-specific expression.
2a	AC-element	ACCAACC ccACCtACCgt AACAAAC	1215 1225 (-) 1268	1447 (-)	Tissue specific and MYB-responsive expression of glutamine synthetase in pine. MYB-dependent activation of the shikimate pathway in response to wounding in Arabidopsis.
2b	E-motifs	ACACxxG	1167 (-) 1818 2389 2448 (-) 2155 (-)	886 (-)	Found in the carrot Dc3 gene promoter that act with TCGTGT distal motifs in response to drought and ABA.
2b*	TCGTGT-motif	TCGTGT	338 1165 1274 (-)	240 (-) 1369	Required distal to E-motifs.
3a	DRE/CRT-core motif	A/GCCGac	1244 1520 1531 1722 1734 1751 1797 1970 1998 2100	2217 (-) 617 962 1091 2030 992 1543 1746 1860	Activates transcription in response to ABA in Arabidopsis. CRT is a C-repeat motif,

			2138 2217 2319 2337 2351 2380 1260 (-) 1435 (-) 1636 (-) 1664 (-) 1905 (-) 2049 (-) 2191 (-) 2203 (-) 2213 (-)	2122	also responsive to drought, high salt and low temperatures.
3b	Em1a Em1b Em2 Like Hex-motif	ACGTGgc ACGTGcc CGAGCAg TGACGT	1171 1299 1841 1874 (-) 1933 (-) 2234 2348 2369 (-)	797 1020 (-)	Activates transcription of Em in response to ABA in wheat. Associated with seed maturation Bound by TGA-1 transcription factor, a bZIP homologue.
3c	ABRE-1 ABRE-2 ABRE-4 Motif II	GACGTG CACGTC CACGTA cCGcCGCGCc	1837 (-) 1837 (-) 1170 (-) 2189 2217 2361		Maize Rab17 promoter ABA-responsive elements. Activation of rice Rab16. ABA-responsive
3c*	ABRE-CE	CATGCCGCC CACCG	329 (-) 1243 1796 1969 2336		Required with ABRE for ABA gene responsiveness
4a	MYB-responsive elements	CAGTTG	1185 1438 (-) 1645 (-)	960	MYB-binding, in tobacco and cotton trichome development
4b	Rice MYB5-responsive elements	AACCAA AACGCA AACAAA AACTAA AACTTA AACAGA AACGTA AACTCA AACCGA AACTGA AACACA AACGAA	1011 (-) 1223 (-) 1337 1467 (-) 1268 1332 1489 (-) 1494 (-) 1404 1470 (-) 1496 (-)	207 933 (-) 1075 (-) 467 (-) 1088 (-) 1320 1199 (-) 1413 (-) 744 1258 526 (-) 928 (-) 969 1023 (-) 2204 (-) 2300	MYB-binding site (reverse complement of GT-box) of AREs identified in maize, Arabidopsis and rice.
4c	CACATG-element	CACATG	2390	885 (-)	Myc-responsive in Arabidopsis in conjunction with Myb-binding element. Responsive to dehydration and ABA.

5	G-boxes	cACGTg	1091 1171 1299 1838	1015	Activation of chalcone synthase in pea epicotyl and parsley (light-responsive). G-box core identified in light-responsive gene promoters. Bound by rice/pea root and epicotyl-specific transcription factors, eg bZIP. Promoter of cereal seed storage proteins.
		GACGTC		798	
		TACCGTA		722	
		GGTTT	1355 1518 1771 (-)		
6	S box	TTTAA	1357 1468 1515	344 382 1022 1469	Closely related to G-boxes. Sugar and ABA-responsive to inhibit transcription.
7	RY-factor	CATGCAt	1683 (-) 2190 (-)		Seed-specific promoter, bound by ABI3 (ABA insensitive), usually -100.
8	Rav1 binding elements	CAACA	1267 1331 1364 1860		Elements both bound by Rav1 (like ABA insensitive- transcription factors) in Arabidopsis, promoting expression in response to abiotic and other stress.
		CACCGT	1614		
9	Dof-binding pyrimidine motif	CCTTTT	1328 1452 1486 1505 1552 1772	733	Bound by plant-specific Dof family transcription factors.
		ACTTTA	1419 1425 1595 1607 1756	823 (-)	
		AAAG		212 679 767 785 843 988 1231 1288 1458 1540 1723 1727 1836	
10	Box D element	AAAGAAAG		482 (-) 1720	Maize Dof1-binding element of cytosolic PPDK and PEP carboxylase promoters.
11	BS element	GNGGTG	1161 (-) 1240 (-) 1445 1969 (-) 2028 (-) 2047 2390 2398 (-)	1880 (-) 1920 (-)	Bound by root-specific transcription factor Alfin1, associated with salt tolerance.
12a	CrichQ	CCCCTCCTC		1953	Promoters of the tomato rbcS2 gene

12b	Box II	GCCACA	2207 2386		Tomato – rbcS3A gene
12c	16KK	CCGTTA	2424 (-)		Tomato – rbcS3A gene
13	Opaque-2 binding element	CACGT	1175 (-) 1303 (-) 1837		Element bound by a bZIP transcription factor, which when mutated affects zein storage in maize endosperm (seeds).
14	CGACG-element	CGACG	1128 1274 2206 (-) 2357	804 972	Present in the promoter region of alpha-amylase, involved with seedling development.
15	SEF-4	gcaTTTTTatca	1465 1771	397 (-) 415 (-) 459 684 (-) 768 (-) 832 1267 1630 1838 (-)	Promoter element found in a soybean seed storage protein gene (SEF – soybean embryo factor).
16a	Box A element	CCGTCC		1679 (-) 1861 1949	Found in promoters of genes involved in flavonoid and phenylpropanoid synthesis pathways. Plant defense response.
16b	Ocs- element	tgaTGACGac		681 782 794 926 938 1890 (-) 1902 (-)	Found in the promoter region of PR genes. SA-responsive.
16c	PR-box	CCGCCG	2388 2366 2222	992	PR gene promoter.
16d	W box	tTTGACT		1704 (-)	WRKY-binding, represses GA-responsive alpha-amylase expression in aleurone cells.

A comparison of sequences up to 1,000 bp upstream of translation initiation in maize and rice was performed for the CA2 gene. Repetitive motifs and patterns were identified, and the highest sequence similarity was found near to the start of translation for both genes (Fig. 4.10).

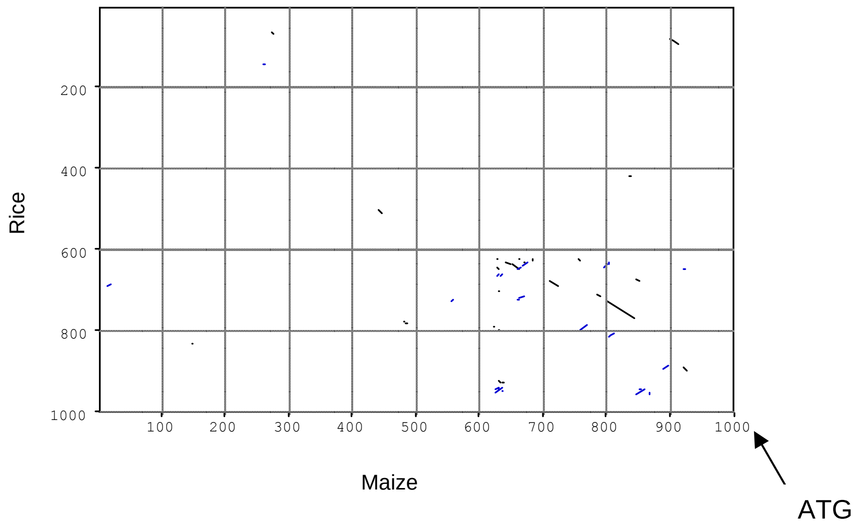


Fig. 4.10. *Comparison of the maize and rice genomic DNA sequences representing 1,000 bp upstream of the translation initiation codon, ATG, of the CA gene for both species. Similarity is based on a minimum similarity of 50% in a 20 bp window (hash value 6).*

In rice, the CA gene codes for a chloroplastic CA, with a high level of homology to the maize CA amino acid sequence (Fig. 4.11).

```
gi606810_maize      MDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVCPSVTLGLQPGE
gi606817_rice       MDAAVDRLKDGFAKFKTEFYDKKPELFEPLKAGQAPKYMVFSCADSRVCPSVTMGLEPGE
                    **.:*:***.** *****.************:**:*:****:*:*********:**:***

gi606810_maize      AFTVRNIASMVPPYDKIKYAGTGSAIEYAVCALKVQVIVVIGHSCCGGIRALLSLKDGAP
gi606817_rice       AFTVRNIANMVPAYCKIKHAGVGSAIEYAVCALKVELIVVIGHSRCGGIKALLSLKDGAP
                    ********.***.* ***:**.*************::******* ****:**********

gi606810_maize      DNFTFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEGLAG
gi606817_rice       DSFHFVEDWVRTGFPAKKKVQTEHASLPFDDQCAILEKEAVNQSLENLKTYPFVKEGIAN
                    *.* ******* * ***:**:.****:******:******** **:***:*******:*.

gi606810_maize      GTLKLVGAHSHFVKGQFVTWEP
gi606817_rice       GTLKLVGGHYDFVSGNLDLWEP
                    *******.* .**.*::  ***
```

Fig 4.11. *ClustalW alignment of the deduced amino acid sequence of Repeat A of the maize CA2 cDNA sequence and the rice chloroplastic CA (not including the chloroplast transit peptide).*

4.3.4 Semi-quantitative RT-PCR

The relative abundance of CA transcripts was analyzed in both root and leaf maize tissues (Fig. 4.12). The relative abundance was determined by semi-quantitative RT-PCR (Fig. 4.13-4.18). The abundance of each transcript was scored based on at which cycle PCR product could be visualized (Table 4.6).

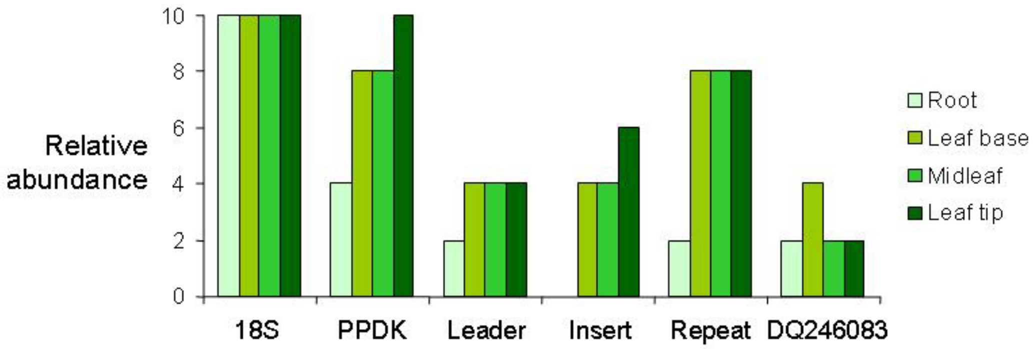


Fig. 4.12. *The relative transcript abundance as determined by semi-quantitative RT-PCR (Fig. 4.13-4.18) using maize roots, and leaf base, middle and tip as sources of template RNA.*

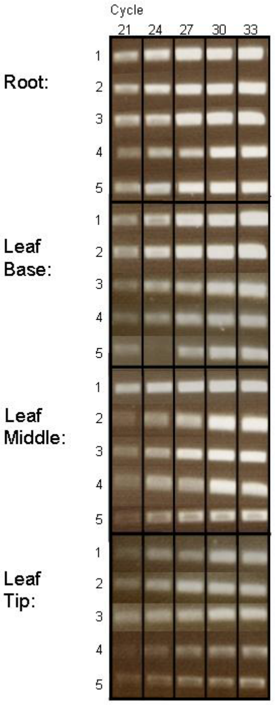


Fig. 4.13. *The results of RT-PCR amplifying a region of the 18S rRNA gene using maize root, leaf base, middle and tips as the source of template RNA. RNA was extracted and cDNA synthesized five times from each tissue source. Reactions were terminated after 21, 24, 27, 30 and 33 cycles as indicated.*

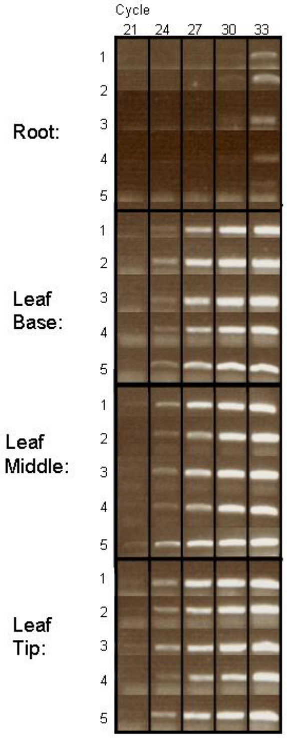


Fig. 4.14. *The results of RT-PCR amplifying a region of the PPDK gene using maize root, leaf base, middle and tips as the source of template RNA. RNA was extracted and cDNA synthesized five times from each tissue source. Reactions were terminated after 21, 24, 27, 30 and 33 cycles as indicated.*

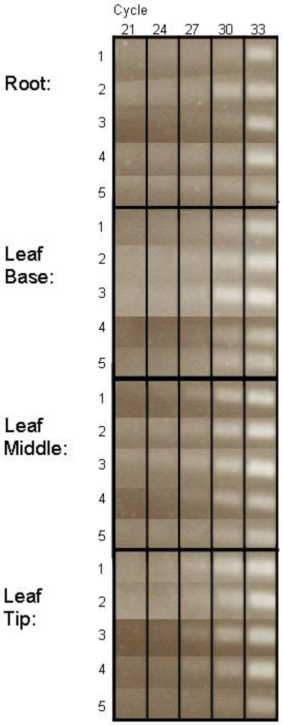


Fig. 4.15. *The results of RT-PCR amplifying a region of the 5'-leader sequence of the CA gene using maize root, leaf base, middle and tips as the source of template RNA. RNA was extracted and cDNA synthesized five times from each tissue source. Reactions were terminated after 21, 24, 27, 30 and 33 cycles as indicated.*

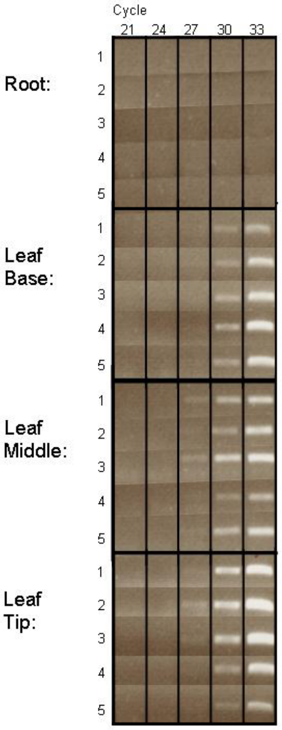


Fig. 4.16. *The results of RT-PCR amplifying a region of the 276 bp insert unique to the CA2 transcript sequence using maize root, leaf base, middle and tips as the source of template RNA. RNA was extracted and cDNA synthesized five times from each tissue source. Reactions were terminated after 21, 24, 27, 30 and 33 cycles as indicated.*

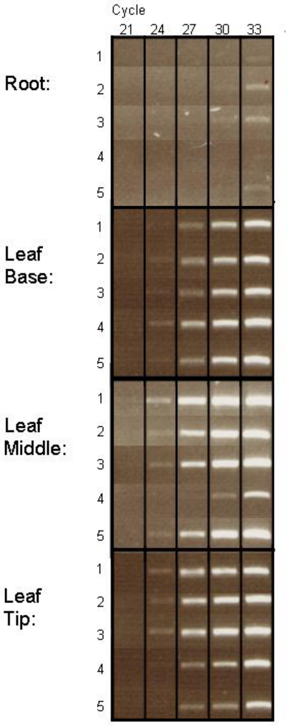


Fig. 4.17. *The results of RT-PCR amplifying part of the repeating regions (Repeat A, B and C) of the CA transcripts using maize root, leaf base, middle and tips as the source of template RNA. RNA was extracted and cDNA synthesized five times from each tissue source. Reactions were terminated after 21, 24, 27, 30 and 33 cycles as indicated.*

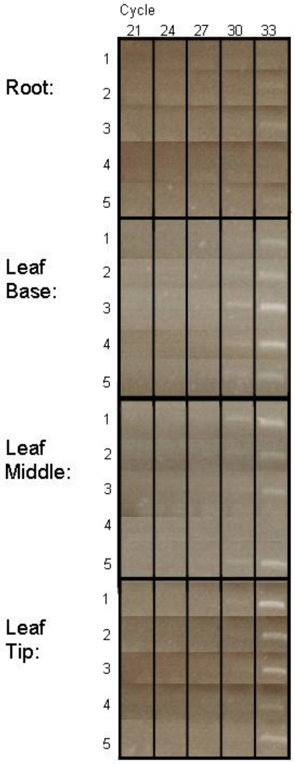


Fig. 4.18. *The results of RT-PCR amplifying a region of the DQ246083 gene using maize root, leaf base, middle and tips as the source of template RNA. RNA was extracted and cDNA synthesized five times from each tissue source. Reactions were terminated after 21, 24, 27, 30 and 33 cycles as indicated.*

Discussion

CAs from the beta-CA gene family have been identified and described in many plant and algal species. However, the CAs of NADP-ME type C_4 monocot species such as maize, sorghum and sugarcane remain relatively uncharacterized. As CA plays an important role in the C_4 photosynthetic pathway by providing the substrate for the initial carboxylation reaction catalyzed by PEP carboxylase, analysis of the primary structure was undertaken. The maize CA sequence was of particular interest as it differed from the sequences of other monocot plant CAs (Fig. 4.5; Burnell and Ludwig, 1997; Wyrich *et al.,* 1998).

Maize is the only C_4 monocot species for which both cDNA and genomic DNA sequences are available to enable gene structure analysis. However, with the completion of the rice genome sequence, the CA gene from this species could be used for comparison (International Rice Genome Sequencing Project, 2005). Rice is also a monocot species, although it is classified as a C_3 plant. In plants that utilize the C_3 pathway, CA has a different function and is located in different sub-cellular compartments compared to C_4 plants. Nevertheless, the structure of the CA genes from maize and rice were compared there are general characteristics that are conserved between the maize and rice CA genes, such as the length of the intron between the first and second exons, and the pattern of two small exons being separated from the adjacent four exons by a relatively large intron (Fig. 4.6). However, the maize CA gene contains two series of six exons, which encode repeating protein domains, while the rice CA gene has only one series of six exons.

Despite the difference in the photosynthetic mechanism used and the function of CA in these plants, the gene structures appear highly conserved. Both maize and rice are members of the grass family Poaceae and have physiological similarities. The duplication of the maize CA gene may be a result of the action of transposons, which are highly represented in the maize genome (Chapter 3).

4.4.1 Analysis of genomic sequence data

4.4.1.1 Intron/exon boundaries

Splicing of introns from pre-mRNA is an essential process for protein expression. Splicing occurs at the spliceosome where, in two trans-esterification steps, exons are ligated and introns are excised. The spliceosome is an assembly composed of small nuclear ribonucleoproteins (snRNPs) and the substrate RNA (Goodall and Filipowicz, 1991). The main factor influencing accurate splicing is the nucleotide sequence across the splice junction at the intron-exon boundary. The nucleotide sequence of the 5′-splice site of the pre-mRNA must be complementary to the RNA sequence of the snRNP U1 (Golovkin and Reddy, 1996).

Using alignments of genomic and cDNA sequence, the exon-intron boundaries were identified (Table 4.7a and b). Predominantly, canonical sequences (*gt-ag)* are present at the exon-intron borders, and these sequences are similar to those found at the boundaries of exons and introns not only in plants, but also yeast and animal genes (Hanley and Schuler, 1988). The consensus sequences were identified in both the CA2 gene sequence and the assemblies obtained from interrogation of the databases (AZM4_68973/4 and AZM4_23203, www.maizegenome.org).

The CA2 gene contains two protein encoding domains, and the sequences over the exon-intron boundaries within these domains are conserved, further emphasizing the homology between the two regions of the gene. The sequences at the 3′-end of the exons also demonstrate a general pattern over two repeats of five exons (*ga, ag, ag, tt, ag*; Table 4.7a). The first three nucleotides of the 5′-splice sequence of the 13 introns were consistent with consensus sequence (*gta)* until the +4 position, at which all four nucleotides were represented. The exception to this was the sequence surrounding Exon 2. The 3′-intron splice sequence was (*tgct*), while the 5′-splice sequence of the second intron was (c*gtc)*.

Accurate splicing is dependent on the nucleotide sequence surrounding introns/exon splice boundaries (Chen *et al.,* 2007). Inaccurate splicing over this exon may result in decreased levels of transcription of the second exon; however this remains to be experimentally shown. Instead, multiple sequencing of maize leaf mRNA via RT-PCR and cDNA library screening have repeatedly shown the existence of this sequence as part of the CA2 transcript (Section 4.3.4 and Section 6.3.3, Chapter 6). Transcription of the CA2 gene could also produce the shortest CA cDNA identified, CA1, if the second exon is excluded. It may be that inaccurate (or accurate) splicing at the exon-intron boundary of Exon 2 results in the production of the CA1 transcript in maize, as opposed to this transcript being encoded by a separate gene.

4.4.1.2 Nucleotide composition

The requirements for accurate pre-mRNA splicing have been shown to be the sequence at the 5′-end and 3′-end of the introns, but also internal sequence forming a splice signal (McCullough *et al.,* 1996). These factors vary between monocots and dicots, where monocot genes that have been introduced into transgenic dicot plants are either not spliced or introns are skipped during processing (Hanley and Schuler, 1988). One of the key differences between monocot and dicot introns is the nucleotide composition, with dicot introns containing a higher proportion of A/U residues. In both types of plants the percentage of A/U residues is higher in introns than exons, with values of 74% for dicot introns, versus 55% for adjacent exons, and for monocots the introns have 56% A/U residues (Goodall and Filipowicz, 1989). The reason monocot introns are not spliced in dicot plants may be the lower percentage of A/U nucleotides.

The nucleotide composition of the CA2 gene and the CA gene from rice were compared (Fig. 4.7A and B). For the maize CA2 gene, most exons are composed of at least 60% G/C residues (40% A/U), while the introns contain less than 50%. The exception to this was the second intron, the nucleotide sequence of which is composed of more than 55% G/C residues (45% A/U). Accurate splicing of introns from pre-mRNA has been shown to be dependent on nucleotide composition and the presence of A/U residues (McCullough *et al.,* 1996). This is significant considering the sequence of the second exon, which appeared to contain changes in the reading frame that would result in non-expression, however without protein sequence data this cannot be confirmed (Section 6.3.3, Chapter 6).

In both the maize and rice CA genes, the nucleotide sequence before the first exon contained a high proportion of G/C residues. Analysis of this sequence identified a 634 bp

CpG island (-690 to -56 bp) in the maize CA gene (Fig. 4.8). Additionally, conservation of the CpG island between maize and rice was observed with the presence of a similarly located CpG island upstream of the rice CA gene. The location of a CpG island at this position is consistent with analysis of the human genome where CpG islands, which are resistant to methylation, were found associated with the transcription initiation site of all constitutively expressed genes (Larsen *et al.,* 1992; Ashikawa, 2001).

4.4.2 Transcription initiation site

The precise transcription initiation site of the maize CA2 gene was not identified. For other genes involved in the C_4 photosynthetic pathway in maize, the region up to 500 bp upstream of the translation-initiating codon has been analyzed for transcription factor binding sites, proximal promoter regions and transcription initiation sites (Matsuoka, 1990; Sheen, 1991; Buchanan *et al.*, 2004; Lebrun *et al.*, 1987). The consensus sequence over the transcription initiation site had been determined for PPDK, PEP carboxylase and chlorophyll a/b binding protein, all of which, like CA, are expressed at high levels in maize mesophyll cells (Matsuoka, 1990). This consensus sequence was identified 394 bp upstream of the site of translation initiation, creating a relatively large 5′-UTR (Table 4.8). The 5′-UTR has been identified as being significant for mRNA stability in the cell, and may protect mRNA from degradation. Using techniques such as RT-PCR and 5′-RACE, the 394 bp 5′-UTR was not detected, although this may also be due to the high percentage of G/C nucleotides composing this part of the gene sequence preventing reverse transcription.

Alternatively, the transcription initiation site may be proximal to the initiating codon, with the sequence identified by PCR representing the actual transcription initiation site, despite not conforming to the consensus sequence (Fig. 4.9, Table 4.9). In plant gene promoters, conservation of sequence around the TATA box and transcription start site motifs for both monocot and dicot species has been analyzed (Shahmuradov *et al.*, 2003). The most highly conserved residues identified in the transcription initiation consensus for 70 unrelated monocots are the C(+3) and A(+4), though in the sequence identified (CGA*CTCG), only the conserved C(+3) residue was present.

4.4.3 Identification of putative regulatory elements

Promoter elements that influence transcription initiation are usually proximal (several hundred nucleotides around the transcription start site), while enhancer elements that regulate transcription can be distal (thousands) and located within introns. Both enable transcription to be controlled spatially, developmentally and in response to environmental conditions. Generally, the sequence upstream of the transcription initiation site is the location for binding of the transcription complex. Enhancers can be located both at the 5′- and 3′-ends of the gene, as well as within introns, where they can function in either orientation and exert either a positive or negative regulatory effect (Chen *et al.,* 1986).

The genes encoding CA in plant leaves (presumed to have a photosynthetic function) have not been previously analysed for the presence or function of promoter or enhancer elements. Rather, changes in transcript levels in response to conditions such as external carbon dioxide partial pressure have been analyzed. For example in pea, a C_3 dicot, expression of CA was modulated by external carbon dioxide levels confirming the role of CA in carbon assimilation (Majeau *et al.*, 1994). Additionally there appears to be coordinated regulation of CA and Rubisco expression in C_3 plant species implying a photosynthetic role for CA.

In marine diatoms, the function of CA is associated with the carbon concentrating mechanism (CCM), concentrating carbon dioxide for photosynthesis. This is supported by activation of CA gene transcription in response to low atmospheric carbon dioxide levels as well as up-regulation in response to light, necessary for the light-dependent reactions of photosynthesis to occur (Harada *et al.*, 2005). In *Chlamydomonas reinhardtii* the transcription of CA is responsive to external carbon dioxide as well as light and the region upstream of transcription initiation of the CA genes (*Cah1 and 2)* in *C. reinhardtii* have been extensively analyzed (Kucho *et al.*, 2003). An enhancer element (GANTTNC) was identified as being essential for transcription in response to low atmospheric carbon dioxide conditions, and this element was also identified in the maize CA2 gene (Table 4.10, Appendix 4). As CA is responsible for converting carbon dioxide that has diffused into the mesophyll cells to bicarbonate for use by PEP carboxylase, more CA would be required in conditions of low atmospheric carbon dioxide concentrations promoting increased gene expression as a compensatory mechanism.

A sequence that is conserved in promoter elements among many eukaryotic genes is the TATA box, which was identified at several positions in the CA2 gene sequence, including the first intron (Appendix 4). Another element that has been identified upstream of transcription initiation in eukaryotes is the CCAAT-box, which is a sequence bound by the CCAAT-binding protein complex. This complex is involved in histone protein acetylation, which then assists gene activation and may even be required for light-responsive gene expression (Benhamed *et al.*, 2006). In an analysis of 131 unrelated plant promoters it was found to have a mean distance from the transcription start site of 75 bp (Shahmuradov *et al.*, 2003), and an inverted CCAAT-box (ATAGG) is found 68 bp upstream of transcription initiation for the PPDK gene (Matsuoka, 1990). There are several potential CCAAT-box motifs found in the CA gene sequence upstream of the putative TATA boxes. There are also several CCAAT-boxes in the 5′-region of Intron 1, but the 3′-region contains predominantly inverted motifs (Table 4.10, Appendix 4).

The anaerobic responsive element (ARE) in the maize alcohol dehydrogenase 1 (*Adh1*) gene is responsible for hypoxic-induced transcriptional activation in a plant tissue-specific manner. The expression of *Adh1* activates ethanolic fermentation pathways in response to hypoxic conditions such as soil flooding. This enhancer functions in both orientations and at variable distances from the TATA box, but is required proximal to the transcription start site for tissue-specific activation (Kyozuka *et al.*, 1994; Dolferus *et al.*, 1994; Olive *et al.*, 1991). AREs have been identified in maize *Adh1, Adh2* and adolase genes, Arabidopsis *Adh*, lactate dehydrogenase and pyruvate decarboxylase, as well as in the CA2 gene sequence both upstream of the start codon and in the first intron (Table 4.10, Appendix 4; Mohanty *et al.*, 2005).

AC element-rich regions are functional as *cis*-acting elements. In carrot (*Daucus carota* L.) the gene *Dc3* is expressed in developing seeds and in vegetative tissue in response to drought and abscisic acid (ABA) treatment. The proximal promoter of this gene has several E-motifs that are characterized as being AC-rich, and which act with TCGTGT-motifs in the distal promoter to activate transcription (Chung *et al.*, 2005). Several of these E-motifs were identified in the CA2 gene sequence (Table 4.10, Appendix 4). Other AC-rich motifs have been reported in the promoters of glutamine synthetase in Scots pine (*Pinus sylvestris)* as well as several species of angiosperms. Several genes for glutamine synthetase exist in Scots pine, which are activated in a tissue-specific and Myb-responsive manner (Gómez-Maldonado *et al.*, 2004). In crop species such as maize, glutamine synthetase is encoded by several genes and has functionally distinct roles based on localization of the enzyme in either the cytosol or the chloroplast, and the role of the promoter region in this localization and function is significant (Miflin and Habash, 2002; Edwards *et al.*, 1990).

The expression of CA in the cytosol or chloroplast, considering the identification of a putative chloroplast transit peptide within the CA amino acid sequence (Section 6.3.1.1, Chapter 6) may also be dependent on the presence of the AC-rich *cis*-element. The promoter region of a plant gene involved in either drought or high salinity stress will have two *cis*-elements, the dehydration-responsive element (DRE; TACCGACAT) and the ABA-responsive element (ABRE; ACGTGG/TC) within 300 bp of the transcription start site (Mukherjee *et al.*, 2006). Dehydration results in the production of ABA, which in turn activates transcription. The same motifs identified in the promoter region of a well-characterized ABA-responsive gene in maize, sorghum and rice was also present in the CA2 gene (Table 4.10, Appendix 4; Buchanan *et al.*, 2004).

G-box elements are *cis*-acting DNA sequences involved in transcriptional activation in response to a variety of stimulants such as light, ABA, and anaerobic conditions. There are many of these elements in the CA2 gene sequence, some of which contain the ACGT-core (Table 4.10, Appendix 4; Hartmann *et al.*, 2005). These include Myb-binding light-responsive elements that are present in the promoters of flavonol synthesis pathway enzymes and the monocot-specific light-responsive GT-box (Seki *et al.,* 1999; Hartmann *et al.,* 2005; Harter *et al.*, 1994). Coordinated regulation of enzymes that respond to light conditions, water availability and carbohydrate production is vital (Rook *et al.*, 2006). Light-responsive genes are often associated with the photosynthetic pathway in plants, with histone acetylation involved in the activation of these genes (Benhamed *et al.*, 2006). The promoter regions of the genes encoding the small subunit of Rubisco, rbcS, have been well studied. In tomato, a dicot species, the promoter region has several conserved elements that are also present in the CA2 gene sequence (Table 4.10, Appendix 4). Generally rbcS expression is light regulated and tissue-specific, as with other enzymes involved in the photosynthetic pathway (Ueda *et al.*, 1989). Light specifically induces rbcS mRNA accumulation in the mesophyll cells of tomato (Kyosuko *et al.*, 1993). However, that CA enzyme activity increases in response to light has been established in species such as cotton and maize, while no corresponding increase in transcript levels was observed (Hoang and Chapman, 2002; Burnell *et al.,* 1990b).

One of the most commonly found motifs in the region of the CA2 gene is the Dof-binding pyrimidine box (Table 4.10). The Dof (DNA-binding with One Finger) family of transcription factors, which are unique to plants, bind these elements usually within 300 bp of the transcription start site. The activity of the Dof proteins has been studied in relation to the activation of several enzymes involved in the C_4 photosynthetic pathway, as well as for other enzymes involved in diverse biological processes such as ascorbate oxidase and a rice peptidase (Yanagisawa, 2000; Yanagisawa, 2004). The expression of enzymes such as PPDK and PEP carboxylase is under tight tissue-specific and light-dependent control, which is due to regulation by Dof transcription factors. Because the Dof transcription factors regulate the expression of a number of genes involved in a particular metabolic pathway, they have also been used to create transgenic plants, improving nitrogen assimilation and amino acid production (Yanagisawa *et al.*, 2004). However, not all promoters containing the conserved AAAG-motif are targets of Dof activation, and for activation of the CA2 gene, which contains several Dof-binding pyrimidine boxes it remains to be experimentally demonstrated (Table 4.10, Appendix 4).

Salicylic acid (SA) plays a role in signaling defense responses in plants, and there are many genes that have been identified from a number of species that are SA-responsive. Glutathione *S*-transferases (GSTs) are involved in the plant stress response and the promoter region of this gene in Arabidopsis was responsive to SA and hydrogen peroxide, predominantly in the roots due to the presence of the octopine synthase (ocs) element (TGACG), first identified in *Agrobacterium* (Chen and Singh, 1999). This element is also found in Intron 1 of the CA2 gene (Table 4.10; Appendix 4). SA is also a necessary signal for systemic acquired resistance (SAR) in plants, a general plant-resistance response that can be induced during a local infection by a pathogen. Accumulation of SA results in the expression of a group of genes known as pathogenesis-related (*PR*) genes (Zhang *et al.*, 1999). In Arabidopsis, a factor known as NPR1 interacts with bZIP transcription factors, for example AHBP-1b, in response to SA to activate transcription of the *PR* genes. These *PR* genes often contain a *cis*-acting element in the gene sequence called a PR-box (GCCGCC), also present in the CA2 gene, and which is the binding site of the tobacco ethylene-responsive element binding proteins (EREBPs) and their homologues (Table 4.10; Zhou *et al.*, 1998).

The SA-binding protein (SABP) 3 of tobacco was identified as a chloroplastic CA (Slaymaker *et al.*, 2002). SABP3 was purified from the soluble fraction of tobacco leaf chloroplasts as a 25 kDa protein that bound SA with an apparent dissociation constant (K_d) of 3.7 μM and silencing of this gene suppressed the hypersensitive response. The promoter of the CA2 gene has elements that direct expression to the chloroplast and a chloroplast transit peptide has been identified in the amino acid sequence (Section 6.3.1.1, Chapter 6). This could signify that CA in the chloroplasts of C_4 plants plays a role in the hypersensitive response, rather than a photosynthetic role.

In order to substantiate the relevance of the promoter-binding elements identified in the maize CA gene, a comparison was made with the rice genomic sequence. Rice is also a monocot species but unlike maize does not use the C_4 photosynthetic pathway; however the rice genomic sequence is annotated and readily accessible from a number of public databases. A comparison of sequences up to 1,000 bp upstream of translation initiation was performed (Fig. 4.10), identifying repetitive motifs and patterns, which showed that most sequence similarity was near the start of translation for both genes. A comparison of the promoter region of the CA2 gene with the sequence of genes involved in similar cellular processes could also provide insight into regulation of CA expression in maize. However, PPDK and PEP carboxylase, which are both involved in the C_4 photosynthetic pathway, do not share similar nuclear-protein binding sequences, despite being regulated in a similar manner (Kano-Murakami *et al.*, 1991). Once the function and location of CA in maize is resolved, this comparison could be more informative.

4.4.4 Analysis of transcript abundance in maize tissues

In order to determine the location of the CA transcripts in maize, the relative transcript abundance was determined and compared. RNA was purified from maize roots, and from three sections along the length of the maize leaf including the base, middle and tip. From this RNA, cDNA was generated and used as template in five replicates of semi-quantitative PCR. Six different primer combinations representing independent reactions were performed (Fig. 4.12), with the relative abundance of the gene-specific reaction products being compared to the amplification of 18s rRNA (Massonneau *et al.,* 2004). In addition to the three reactions amplifying the CA2 transcript, amplification of a 545 bp region of the putative CA transcript identified in the NCBI database (DQ246083, http://www.ncbi.nlm.nih.gov) was performed to further characterize this CA transcript.

Three different regions of the CA transcript were amplified, the 5′-leader sequence (a putative chloroplast transit peptide, Section 6.3.1.1, Chapter 6), the 276 bp insert unique to CA2, and the repeating region of the transcript (Fig. 4.4). The amplification of each component could then be compared in order to determine whether the transcript contained either the 5′-leader sequence, or the 276 bp unique insert, both, or neither, in association with the repeating regions (Repeat A, B and C), which could not be distinguished due to over 90% sequence homology (Section 4.1.3). There is no protein sequence data from leaf (or root) purified CA, and as the transcript lengths (CA1, CA2 and CA3) do not correlate to protein subunit size obtained by western blotting of leaf extracts, this implies that some form of modification must occur (Burnell and Ludwig, 1997). By amplifying the different regions of the CA transcript independently, the aim was to determine whether post-transcriptional or post-translational modification was occurring.

The purpose of dividing the leaf into three sections was to obtain RNA from cells at different stages of development. Cell division occurs primarily from the basal meristem of the leaf, which results in the youngest cells being present at the base of the leaf, and the older cells at the leaf tip (Martineau and Taylor, 1985; Martineau and Taylor, 1986). Cellular differentiation, both morphological and functional, also occurs as the cells mature resulting in the production of bundle sheath and mesophyll cells necessary for operation of the C_4 photosynthetic pathway. Previous studies have shown that it is not until this differentiation occurs that C_4-specific transcripts are detectable (Langdale *et al.,* 1988a; Langdale *et al.,* 1988b). Therefore, using the leaf base as a source of RNA was hypothesized to provide young undifferentiated cells that would be reliant on a single-celled C_3 pathway being the predominant means of photosynthesis, and hence a chloroplastic location of CA transcripts was expected. In contrast, the transcripts amplified from the leaf tip would be representative of fully differentiated, mature bundle sheath and mesophyll cells operating using the C_4 pathway. In this way the function of CA in the leaf might be predicted.

Amplification of a region of the well-characterized C_4-specific transcript PPDK was used as a positive control to show whether the RNA isolated from along the length of the maize leaf was actually representative of different stages of cellular differentiation and development. Like other C_4 enzymes, it was expected that the PPDK transcript would increase in abundance from the leaf base to the leaf tip (Langdale *et al.,* 1988a). Amplification of a region of the PPDK transcript using RNA purified from along the length of the maize leaf did not show the expected abundance gradient from leaf base to leaf tip (Fig. 4.12). This indicated that the RNA purified from the base and middle sections of the leaf was most likely from cells differentiating into mesophyll and bundle sheath cells and operating using the C_4 pathway. However, the relative abundance was greater at the leaf tip, and amplification of the 276 bp insert region of CA2 also showed an increase in relative abundance towards the leaf tip (Fig. 4.12). From this observation, it could be concluded that this part of the transcript is associated with the CA isoform involved in the operation of the C_4 pathway. Hence, this CA isoform would be located in the cytosol of mesophyll cells, and may be the membrane-associated 28 kDa isoform (Burnell and Ludwig, 1997; Utsunomiya and Muto, 1993). The deduced amino acid sequence of the 276 bp insert is hydrophilic, and therefore not likely to be located in the membrane. However, it has been previously hypothesized that it may be associated with a membrane protein (Burnell and Ludwig, 1997).

Amplification of a region of the 5′-leader sequence (the putative chloroplast transit peptide) indicated that this component of the CA transcript was being produced at relatively low levels along the full length of the maize leaf (Fig. 4.12). As the activity of CA in the bundle sheath cell chloroplast would be detrimental to the operation of the C_4 pathway (Burnell and Hatch, 1988), the identification of a putative chloroplast transit peptide as part of the CA transcript was unexpected (Section 6.3.1.1, Chapter 6). However, comparing the relative abundance of this transcript component with the relative abundance of the repeating regions of the transcript indicated that the putative chloroplast transit peptide was not always transcribed along with the Repeats, and rather may only be associated with one of the repeating regions (Repeat A, Repeat B or Repeat C).

In *F. bidentis,* the chloroplast transit peptide is retained on all three isozymes though does not promote transport into the chloroplast, as only one of the CA isoforms identified in that species was imported into chloroplasts in *in vivo* uptake assays (Tetu *et al.,* 2007). It should be noted that the reaction used to amplify the repeating regions used primers that did not distinguish between Repeat A, B and C, nor between CA1, CA2 nor CA3, and so the transcript abundance might be a reflection of amplification of all three Repeats from all three transcripts. Nevertheless, if CA transcripts that included the 5′-leader sequence had been higher at the leaf base, where functionally undifferentiated cells were using the C_3 photosynthetic pathway to generate energy, this would explain the need for a chloroplast transit peptide on a protein, the activity of which would be otherwise detrimental to the operation of the CCM. However, it is unlikely a true representation of these undifferentiated cells was obtained in this experiment, as observed by the PPDK amplification results (Fig. 4.12).

In the inner cortex of legume nodules of species such as alfalfa (*Medicago sativa*) and soybean (*Glycine max*) the first non-photosynthetic CAs identified had anaplerotic roles in replenishing the Krebs cycle intermediates together with PEP carboxylase (Gàlvez *et al.,* 2000; Kavroulakis *et al.*, 2000). The expression levels of the alfalfa CA were responsive to oxygen conditions, enabling coordinated regulation of carbon and nitrogen metabolic pathways. CA transcripts were identified in maize root tissue (Fig. 4.12), and have also been identified in root tissue of *F. bidentis* (Tetu *et al.,* 2007). Although expressed at much higher levels in leaf tissue the presence of both the repeating region of the CA transcript and the DQ246083 transcript in root tissue may indicate additional non-photosynthetic roles for CA in maize.

Most genes that encode proteins that play a role in the C_4 pathway are part of a gene family and contain members which have non-photosynthetic functions and are constitutively expressed in a wide range of tissues. For example, the PEP carboxylase multi-gene family contains members which have photosynthetic and non-photosynthetic functions. The non-photosynthetic functions of PEP carboxylase include replenishment of tricarboxylic acid cycle intermediates and carbon skeletons for amino acid biosynthesis (Chollet *et al.,* 1996). It may be that the non-photosynthetic forms of PEP carboxylase also require a supply of bicarbonate for these processes, provided by CA. In leaves, PEP carboxylase provides the initial carboxylation reaction of the C_4 pathway using bicarbonate to produce four carbon acids such as malate. The mesophyll cell-specific expression of photosynthetic PEP carboxylase was found to be dependent on *cis*-elements in the promoter of the gene (Ewing *et al.,* 1998). CA expression for both non-photosynthetic and photosynthetic functions may have evolved by an analogous mechanism.

Conclusion

With the availability of both genomic and cDNA sequence, the CA2 gene of maize could be analyzed. The CA isozymes from maize are unique from other plant species being encoded by comparatively large transcripts, due to the presence of repeating sequences (Burnell and Ludwig, 1997; Wyrich *et al.,* 1998). Three CA transcripts were previously identified that were 2.2, 1.9 and 1.3 kb in length and were found to be composed of two or three repeating sequences of approximately 600 bp. These encode for nearly identical protein domains, which were designated Repeat A, Repeat B and Repeat C. The characterization of CA primary structure provided a means to obtain information related to the expression, function and localization of the CA isozymes.

The CA2 gene encoding the 1.9 kb transcript contains a unique 276 bp insert, which corresponded to a single exon (Chapter 3). The primary structure of this insert was analyzed, and it was found that accurate splicing was unlikely due to the presence of non-canonical intron/exon boundaries. Furthermore, the nucleotide composition of this exon did not conform to that observed for the remaining exons composing the CA2 gene, nor for exons composing the orthologous rice CA gene. While this component of the CA transcript was detectable by reverse transcriptase PCR, indicating that transcription was occurring, analysis of transcript abundance by semi-quantitative PCR indicated that this was at a lower level than transcription of the repeating regions of the CA2 gene. Whether this unique part of the CA transcript is translated to form part of the expressed protein in the maize leaf remains to be determined.

Analysis of the CA2 gene promoter indicated that elements that promote both constitutive and cell-specific or light-responsive expression were present. In addition, a CpG island was identified upstream of the translation-initiation site. Detection of CA transcripts in root tissue, and the presence of multiple isoforms confirm a non-photosynthetic function for CA in maize. The photosynthetic role of CA was supported by the finding of light-responsive elements in the CA2 promoter such as the CCAAT-box and Myb-binding elements, which are often associated with photosynthetic genes (Benhamed *et al.,* 2006).

The function of CA in the chloroplast, however, still remains unclear. Like the gene encoding glutamine synthetase, the CA2 gene promoter also contained several AC-rich *cis*-elements that have been shown to direct expression to different sub-cellular locations (Miflin and Habash, 2002). CA (SABP3) in the chloroplasts of tobacco forms part of the hypersensitive defense response (Slaymaker *et al.,* 2002). Alternatively, there is strong evidence of at least two CAs forming part of Photosystem II in the thylakoid membranes of chloroplasts (Stemler, 1997). One of these CAs may be from the alpha gene family, as an antibody produced against a thylakoid lumen-targeted alpha-CA from *C. reinhardtii* reacted with a 33 kDa polypeptide in the thylakoids of maize mesophyll cells (Lu and Stemler, 2002), and due to antigenic differences between alpha- and beta-CAs, cross-reactivity would be unlikely (Yang *et al.,* 1985). The localization of a CA isozyme in the chloroplast of the C_4 dicot *F. bidentis* was confirmed by *in vivo* import assays and immunohistochemistry, while another isozyme was confirmed to be located in the cytosol despite retaining the chloroplast transit peptide (Tetu *et al.,* 2007). It may be that the putative chloroplast peptide remains after the location of CA changed during evolution from the C_3 to the C_4 pathway (Section 6.4.1.1, Chapter 6).

The abundance of transcript representing the repeating regions (Repeats A, B and C) observed by semi-quantitative PCR was relatively high along the length of the maize leaf. This may be due to amplification of all three Repeats, or could indicate that the Repeats are transcribed independently of the 5′-leader sequence and the 276 bp CA2 insert. Due to over 90% nucleotide sequence homology between the three repeats, quantitative amplification of a single repeat was not possible. Whether the single repeats demonstrate CA activity when expressed has been investigated further (Chapter 5).

CHAPTER 5

5. Expression and Characterization of Maize CA

Introduction

The CA isozymes from maize are unique from other plant species being encoded by comparatively large transcripts (Burnell and Ludwig, 1997; Wyrich *et al.,* 1998). The length of the CA transcripts is due to the presence of repeating sequences. Three CA transcripts were previously identified that were 1.3 kb, 1.9 kb and 2.2 kb and were composed of two or three repeating sequences of approximately 600 bp. These encode nearly identical protein domains, which were designated Repeat A, Repeat B and Repeat C (Burnell and Ludwig, 1997).

5.1.1 Molecular structure of maize CA

Analysis of RNA from maize by northern blot showed three bands of approximately 1.3 kb, 1.9 kb and 2.2 kb that hybridized with a maize CA cDNA probe (Burnell and Ludwig, 1997). The nucleotide sequences of the corresponding cDNAs have been determined and have been designated CA1, CA2 and CA3 respectively (Fig. 5.1; Burnell and Ludwig, 1997; Burnell, 2000). All three CA cDNAs are composed of repeated sequences.

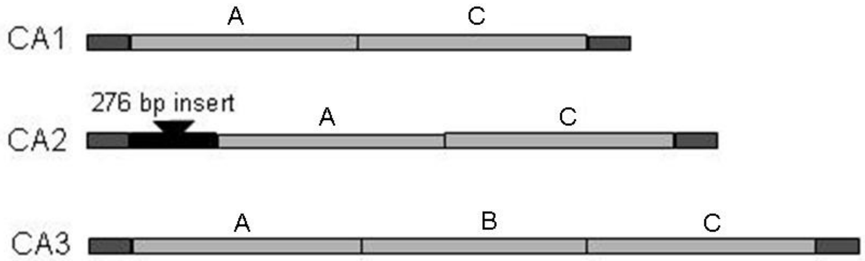


Fig. 5.1. *Schematic representation of the three beta-CA cDNAs from maize. The repeat sequences are approximately 600 bp and are labelled A, B and C. The black shaded box represents the 276 bp insert unique to CA2. The dark grey shaded boxes represent the 5'-leader sequences and 3'-untranslated regions.*

The primary structures of the CA cDNA sequences from maize appear to be specific to NADP-malic enzyme type C_4 monocots (Burnell, 2000). The repeating sequences have been designated Repeat A, Repeat B and Repeat C, and each encodes a discrete protein domain. The longest CA cDNA, CA3, is composed of a 5′-leader sequence as well as Repeat A, Repeat B and Repeat C. CA2 does not contain Repeat B, but consists of Repeats A and C as well as a unique 276 bp insert and the same 5′-leader sequence as CA3. CA1 is similar to CA2, although does not contain the unique 276 bp insert. Both the nucleotide and deduced amino acid sequences of Repeat A and Repeat C in CA1, CA2 and CA3 are identical and are homologous with each other and with Repeat B (Fig. 5.2). Furthermore, it appears that Repeat B may be a chimeric product of Repeat A and Repeat C, with the N-terminus of Repeat B having homology with the N-terminus of Repeat C, while the C-terminus of Repeat B is more similar to the amino acid sequence at the C-terminus of Repeat A.

```
RB             MDPPQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMVFACSDSRVCPSVTLGLQ
RC             MDPPQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMVFACSDSRVCPSVTLGLQ
RA             ---MDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVCPSVTLGLQ
                   : ::***.****:**.:********* ******:*:********************

RB             PGEAFTVRNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIRALLSLKD
RC             PGEAFTVRNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIRALLSLQD
RA             PGEAFTVRNIASMVPPYDKIKYAGTGSAIEYAVCALKVQVIVVIGHSCCGGIRALLSLKD
               ***********:*** *** **:* *************:*:*****************:*

RB             GAPDNFHFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEG
RC             GAPDTFHFVEDWVKIGFIAKMKVKKEHASVPFDDQCSILEKEAVNVSLENLKTYPFVKEG
RA             GAPDNFHFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEG
               ****.********:**  ** ***************************:***:*******

RB             LAGGTLKLVGAHYDFVKGQFVTWEPP
RC             LANGTLKLIGAHYDFVSGEFLTWKK-
RA             LAGGTLKLVGAHYDFVKGQFVTWEPP
               **.*****:*******.*:*:**:
```

Fig. 5.2. *Alignment of the amino acid sequences of Repeat A (RA), Repeat B (RB) and Repeat C (RC).*

5.1.2 Enzymatic activity

CA participates in a range of biological functions including carboxylation reactions, acid-base balance, ion exchange and the cellular processes of respiration as well as photosynthesis, in a wide range of organisms (Tashian, 1992). CA is a zinc metallo-enzyme that catalyses the reversible hydration of carbon dioxide to bicarbonate, significantly increasing the rate of this spontaneously occurring reaction (Fig. 5.3; Burnell, 1990). Generally CA is recognized as catalyzing the first reaction of the C_4 photosynthesis pathway, hydrating carbon dioxide that has diffused into the mesophyll cell cytoplasm to form bicarbonate, the inorganic carbon source for the C_4 pathway (Hatch and Burnell, 1990).

$$CO_2 + H_2O \rightleftharpoons HCO_3^- + H^+$$

Fig. 5.3. *Reaction catalyzed by CA.*

The CA from spinach has a k_{cat} of 2 x 10^5 s^{-1}, with carbon dioxide hydration increasing with increasing pH (Rowlett *et al.,* 1994). The k_{cat} for the hydration activity of rice CA, expressed in and purified from *E. coli,* was 1.1 x 10^5 s^{-1}, and like alpha-CAs was significantly inhibited by sulfonamides, specifically acetazolamide (Pocker and Ng, 1974; Yu *et al.,* 2007; Lehtonen *et al.,* 2004). Sulfonamides inhibit the alpha-CAs by binding to the catalytic zinc molecule (Vidgren *et al.,* 1990). This implies that for beta-CAs from plant species the zinc molecule also has a catalytic role (Bracey *et al.,* 1994; Rowlett *et al.,* 1994). The inhibitors of CA activity have been extensively studied in dicot species such as spinach and pea, where heavy metal anions also decrease CA activity (Atkins *et al.,* 1972b; Ivanov *et al.,* 2007).

CA activity varies depending on the assay buffer due to inhibition of the enzyme by carbon dioxide at the high carbon dioxide concentrations used to ensure substrate saturation, and the ability of the buffer to act as a proton acceptor. Buffers such as imidazole, Tricine, Hepes and Tris usually record lower activities than when CA is assayed in barbitone buffer (Hatch, 1991). The activity of CA is coupled with the production of hydrogen ions (Fig. 5.3), and a colorimetric assay can be used to detect associated changes in pH. The reaction has been well characterized for alpha-CAs, and proceeds by a two stage ping pong mechanism (Fig. 5.4). In the first stage the hydroxide ion, which is highly nucleophilic, attacks the highly electrophilic carbon of the carbon dioxide molecule. The second stage of the reaction is thought to be the rate-limiting step and involves regeneration of the active enzyme state by donation of a proton to the buffer molecule (Rowlett *et al.,* 1994).

$$\mathrm{EZn^{2+}OH^{-} + CO_2 + H_2O \rightleftharpoons EZn^{2+}OH_2 + HCO_3^{-}}$$

$$\mathrm{EZn^{2+}H_2O + B \rightleftharpoons EZn^{2+}OH^{-} + BH^{+}}$$

Fig. 5.4. *The two stage ping pong reaction mechanism described for alpha-CAs (Kimber and Pai, 2000). E represents the enzyme and B represents the buffer molecule.*

5.1.3 Rational

The aim of this chapter was to express and characterize CA and to determine CA activity. Nine expression plasmid constructs were created for this analysis. The repeating regions of the CA transcripts were defined as discrete protein domains, and whether these could independently catalyze the hydration of carbon dioxide was established. The activity of each of the expressed proteins was measured to determine if there were any differences in enzymatic activity despite the similarities in primary structure, and the sensitivity to common CA inhibitors was investigated and compared.

Materials and methods

5.2.1 Source of CA plasmid constructs

CA1 corresponded to the shortest CA transcript (1.3 kb; Burnell and Ludwig, 1997) and contained only Repeat A and Repeat C (Burnell, unpublished), and was provided inserted into an expression vector (pROEx^{TM}-HTb, Invitrogen; Appendix 5.1). The longest transcript (2.2 kb) has since been designated CA3 and contained the 5′-leader sequence, Repeat A, Repeat B and Repeat C (Burnell, 2000). CA2 was provided inserted in the plasmid vector pBluescript II SK(+/-) (Stratagene; Appendix 5.1); however after sequence analysis was found to contain several errors in the reading frame (Section 6.3.3, Chapter 6). Repeat B and Repeat C were also provided as constructs inserted into the expression vector, pROEx^{TM}-HT.

5.2.2 Creation of CA expression constructs

The plasmid constructs supplied for this analysis included CA1, Repeat B and Repeat C, which were provided in the expression vector pROEx^{TM}-HT. The Repeat B construct had an additional 48 nucleotides encoding 16 amino acids at the C-terminus due to cloning, which were not part of the reported maize CA cDNA sequence (Genbank accession U08403, version gi:606814, Appendix 5.2).

Repeat A and the other plasmid constructs used in this analysis were generated by PCR or restriction enzyme modification of existing plasmid constructs. These included a construct containing the 5′-leader sequence, Repeat A and Repeat B (CA4), and several chimeric constructs that did not contain the 5′-leader sequence. Of these the first contained the 5′-end of Repeat C and the 3′-end of Repeat A (CA5), the second contained the 3′-end of Repeat A adjacent to full-length Repeat C (CA6), the third contained the 5′-end of Repeat A and the 3′-end of Repeat C (CA7), and a fourth construct containing only the 3′-end of Repeat A (CA8; Fig. 5.5).

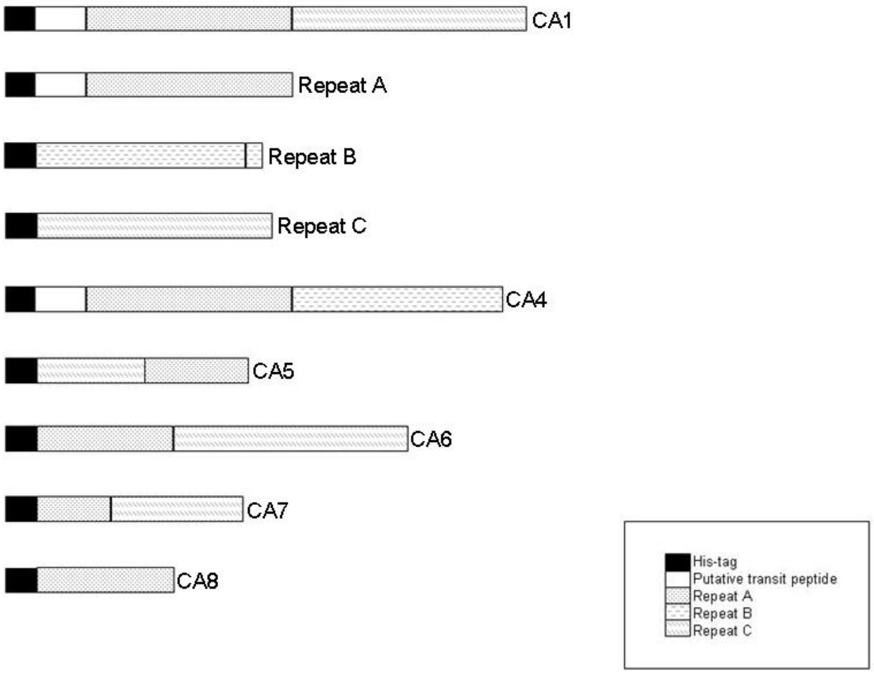


Fig. 5.5. *Schematic representation of the components of the CA transcripts that were expressed and used to analyse CA activity. The histidine tag is represented as a black box and the 5′-leader sequence is represented as a putative transit peptide.*

5.2.2.1 RT-PCR primers and reactions

The clones generated by amplification of cDNA generated from maize leaf mRNA included: CA4, CA7 and CA8. CA4 was amplified using a 5′-primer that bound across the start of the 5′-leader sequence (5′-cacgaggcgcaggatccatgtacacat-3′) and incorporated a *Bam*HI recognition site in the correct reading frame for protein expression using the vector pROEx^{TM}-HT-B. The reverse primer for this reaction was designed to anneal to the 3′-end of Repeat B, however this sequence is also present at the 3′-end of Repeat A (5′-cctagggaggctcccatgtgacg-3′), thereby generating both Repeat A and the CA4 PCR product. The cDNA for the other constructs was amplified using a 5′-primer that hybridized to the sequence at the start of Repeat A (5′-aaggccatggaccccaccgtcg-3′) incorporating an *Nco*I recognition site, and a reverse primer that bound to the unique sequence at the 3′-end of

Repeat C (5′-agccctagtttttcactttttccatg-3′). The cycling conditions used included two rounds of amplification, with a higher annealing temperature initially to ensure amplification of specific gene products (Table 5.1).

Table 5.1. Reaction conditions for amplification of CA from maize leaf mRNA

Cycling Conditions	Step	Temperature	Time
	1	95ºC	5 min
	2	95ºC	45 s
	3	55ºC	45 s
10 cycles from Step 2	4	72ºC	3 min
	5	95ºC	45 s
	6	52ºC	45 s
25 cycles from Step 5	7	72ºC	3 min
	8	72ºC	10 min

The PCR products obtained were sub-cloned using the pGEM®-T vector system (Promega, Appendix 5.1), and the sequence analysed. Subsequently, the CA cDNA inserts were sub-cloned into the expression vector pROEx^{TM}-HT-A, or -B, using the restriction enzyme recognition sites incorporated by the 5′-primers and 3′-restriction sites of the pGEM®-T vector multiple cloning site.

5.2.2.2 Restriction enzyme generation of constructs

Several of the CA cDNA constructs used in this analysis were created by restriction enzyme modification of existing CA plasmid constructs. Repeat A was amplified by PCR using CA2 clone as template, sub-cloned into the pGEM®-T vector, and then sub-cloned into pROEx^{TM}-HT using the *Bam*HI recognition site generated by the 5′-primer sequence and the *Kpn*I recognition site present in the pGEM®-T vector multiple cloning site. The CA6 chimera was generated by amplification using the CA2 clone provided as template and using the same primers as for the reaction used to generate Repeat A. This PCR product was also sub-cloned into the pGEM$^{®}$-T vector, before sub-cloning into pROEx^{TM}-HT using *Nco*I and *Spe*I recognition sites. The restriction enzyme *Nco*I has a recognition site within the Repeat A sequence at position 206, thus generating the chimeric product containing only the 3′-end of Repeat A (Fig 5.6).

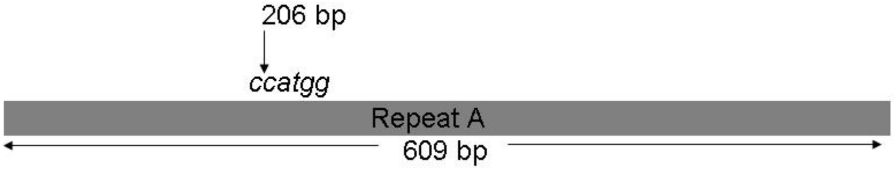


Fig 5.6. *Schematic representation of Repeat A showing the* Nco*I recognition site within the Repeat A sequence.*

CA5 was generated by amplification with the same primer combination used to generate CA6, thereby it also contained only the 3′-end of Repeat A from the *Nco*I site. The construct containing nucleotide sequence encoding only the C-terminus of Repeat A (CA8) was generated with the same 5′-primer, and subsequently sub-cloned into pGEM$^{®}$-T, before sub-cloning into pROEx^{TM}-HT using *Nco*I, and a 3′-restriction enzyme site from the pGEM$^{®}$-T vector multiple cloning site, resulting in the construct containing only sequence corresponding to the 3′-end of Repeat A.

5.2.2.3 Sequence analysis

The DNA sequence of all constructs was confirmed by sequencing from both primer binding sites of the expression vector pROEx^{TM}-HT, which are M13R and pROExRev (5′-tatcaggctgaaaatcttctctc-3′) and from internal sites. The amino acid sequences of the expressed CA enzymes are presented in Appendix 5.2.

5.2.3 Bacterial expression and purification

The CA enzymes were expressed with an N-terminal histidine tag using the pROEx^{TM}-HT vector. The enzymes were expressed in *E. coli* (NM522 cells) and purified by immobilized metal affinity chromatography (IMAC) using HisTrap HP affinity nickel columns (Amersham Biosciences).

5.2.3.1 Inoculating cultures and inducing expression

The plasmid constructs were transformed into chemically competent NM522 cells by heat shock (Section 2.1.9, Chapter 2) and grown on LB agar plates containing carbenicillin/ampicillin (50 µg.ml^{-1}) selection. A single colony was used to inoculate a 5 ml primary LB culture, which was grown overnight at 37ºC. This culture was used to inoculate a 500 ml culture in a 2 L flask and grown with shaking until the optical density of cells was 0.6 at 600 nm. Expression was induced by the addition of 1 mM IPTG, and the cultures were incubated for an additional 12-20 h at 25ºC with agitation.

5.2.3.2 Purification by affinity chromatography

The bacterial cells used for protein expression were harvested by centrifugation at 2,990 × *g* for 10 min. The cell pellets were washed and resuspended in CA buffer (100 mM Tris-HCl pH 7.5, 5 mM MgCl, 200 mM NaCl) in twice the volume of the cell pellet. Cell lysis was facilitated by freezing at -80ºC, before sonication on ice for 12 ten-second bursts. The lysate was cleared by centrifugation at 20,000 rpm for 30 min at 4ºC using a Beckman JA25.5 rotor. A sample of this crude extract was kept for analysis by SDS-PAGE and Western blotting (Section 5.2.4.1) as well as measurements of CA activity (Section 5.2.4.5).

The affinity columns were prepared and charged as described by the manufacturer (Amersham Biosciences). The affinity columns were equilibrated with CA buffer before addition of the cleared lysate that had been filtered through a 0.45 µm filter, and subsequently washed with CA buffer. Protein was eluted from the column with elution buffer, which was the CA buffer with the addition of increasing (5-200 mM) imidazole concentrations. Movement of protein through the column was followed spectrophotometrically at 280 nm using Pharmacia Biotech optical equipment, with a trace printed to a Bio-Rad Model 1325 Econo-Recorder. A Pharmacia Biotech peristaltic pump was used and eluted fractions were collected using a Pharmacia Biotech Frac-100 fraction collector. The concentration of protein in the eluted fractions was calculated using a Bradford Assay (Section 2.2.5, Chapter 2).

5.2.4 Analysis of the expressed CA enzymes

5.2.4.1 SDS-PAGE and Western blots

The samples eluted from the affinity column as well as samples of the crude extract were analysed by SDS-PAGE as described (Section 2.2.1, Chapter 2). Immunoblotting was performed with three different primary antibodies. Anti-CA (α-CA) was generated in rabbit against bacterially-expressed and purified CA1 (Burnell Laboratory, James Cook University, Australia), and was used as a polyclonal serum with 5% (w/v) milk powder blocking agent at a dilution of 1:3,000. The α-CA antibodies were also blot-affinity purified (α-pCA) as described in Section 5.2.4.2. These were used with 5% (w/v) milk powder blocking agent at dilutions up to 1:500. Antibodies against the N-terminal histidine tag were used to confirm expression of the recombinant proteins. Mouse anti-TetraHis (α-His) antibody (QIAGEN) was used with 3% (w/v) BSA blocking agent at a dilution of 1:2,000. When using α-CA, goat anti-rabbit IgG(H+L)-HRP conjugate (BioRad) was used as secondary antibody at 1:1,000 dilution. When using α-His, goat anti-mouse IgG(H+L)-HRP conjugate (Promega) was used as secondary antibody at 1:2,500 dilution.

5.2.4.2 Blot-affinity purification of antibodies

Anti-CA antibodies (α-CA) were purified from rabbit antiserum (Section 5.2.4.1) by blot-affinity purification using Repeat B, that had been bacterially expressed (Section 5.2.3) and purified by affinity chromatography. Repeat B samples (approximately 5 mg total protein) were electrophoresed across a 12.5% (w/v) polyacrylamide mini-gel (BioRad) and the protein transferred to polyvinylidene fluoride (PVDF) membrane. Narrow strips, representing the width of one lane (of a 10-well gel) were cut from either side of the PVDF membrane and blotted using the α-CA antibody as described (Section 5.2.4.1) to locate

Repeat B on the membrane. A 3-4 mm strip across the membrane, corresponding to the location of Repeat B, was excised and placed in a 10 ml Falcon tube. The strip was blocked with 5% (w/v) milk powder in TBST (10mM Tris-HCl pH 7.5, 150 mM NaCl, 0.05% v/v Tween-20) for 30 min before addition of α-CA at a dilution of 1:2,000. The membrane was incubated with the antibody overnight with agitation. The purified antibodies (α-pCA) were eluted from the strip after three 10 min TBST washes with 1 ml of 0.1 M glycine, pH 2.6 at room temperature for 30 min. The elution buffer was then transferred to a 1.5 ml microcentrifuge tube and the pH neutralized by the addition of 50 µl 1.0 M Tris-HCl pH 8.0. The solution was then used as a source of primary antibody at dilutions from 1:1,000 to 1:500.

5.2.4.3 Antibody clearing

In order to decrease the presence of non-specific antibodies in the rabbit antiserum (Section 5.2.4.1), it was incubated with a bacterial extract before use. The bacterial extract was prepared by sonication of a sample of *E. coli* liquid culture, and cleared by centrifugation at 20,000 rpm for 30 min at 4ºC using a Beckman JA25.5 rotor. 0.1 ml of the supernatant was incubated with 5 µl of α-CA for a minimum of 30 min at 4°C before the mixture was added to the membrane in TBST with 5% (w/v) milk powder. The membrane was washed and stained with secondary antibodies as described (Section 2.2.4, Chapter 2).

5.2.4.4 Secondary structure prediction based on the amino acid sequence

The secondary structures of Repeat B and Repeat C were predicted and compared by the use of the protein secondary structure finder at Softberry (www.softberry.com).

5.2.4.5 Measuring CA activity

CA activity was calculated by determining the time required to decrease pH by 0.2 units within the range of pH 8.5 to pH 8.0, measured with a pH electrode. The reaction was performed in 3 ml total volume and contained 2 ml of 25 mM barbitone buffer (pH 8.5) and the protein sample. The time taken to decrease the pH was determined after the addition of 1 ml of carbon dioxide-saturated distilled water. For the uncatalyzed reaction, an equal volume of protein buffer (0.5 M NaCl, 5 mM $MgCl_2$, 50 mM Tris-HCl pH 7.5) was included in the reaction. The rate of the uncatalyzed reaction was then subtracted from the catalyzed rate. The units of activity were determined for the amount of CA added to the reaction, using a Bradford assay (Section 2.2.5, Chapter 2) and determining the proportion of CA in the protein mixture after affinity chromatography by visualization of a Coomassie blue-stained electrophoresed sample (Section 5.2.3.2). Independent reactions were repeated several times to determine the average rate observed, with the average value reported and the standard error calculated.

One unit of enzyme activity was calculated as the µmol of carbon dioxide consumed per second (µmol.s^{-1}), determined by titration of the 25 mM barbitone buffer with hydrochloric acid in order to establish how many µmol of hydrogen ions were required to decrease the pH by 0.2 units in a volume equal to the reaction volume (3 ml). It was determined by titration that a decrease in pH of 0.2 units was the equivalent of the addition of 6 µmol of hydrogen ions, or 6 µmol of carbon dioxide consumed in the reaction catalyzed by CA (Fig. 5.3).

An example of the calculations performed is presented below for Repeat B. After analyzing the purified sample by SDS-PAGE and Coomassie blue staining, it was determined that Repeat B composed 40% of the total protein sample, the concentration of which was 5.4 µg.µl^{-1}. In order to obtain a measurable rate the protein sample was diluted 1:100, and 10 µl was used in the CA assay (0.54 µg), of which 40% was Repeat B (0.22 µg). The amount of carbon dioxide consumed per second was determined for each recorded time measurement, and the uncatalyzed reaction rate was subtracted from this value. The rate was then determined for the amount of protein assayed. The standard deviation was determined for each measurement taken (n = 7), and the standard error determined as the standard deviation divided by the square root of n.

5.2.4.6 CA inhibition analysis

The activity of CA was measured in the presence of four compounds; 1 mM acetazolamide, 1 mM 6-ethoxyzolamide, 10 mM 1,10-phenanthroline and 1% (v/v) diethylpyrocarbonate. For these comparisons, the compound was added to a sample of the enzyme and incubated on ice for 10 min before measurements were taken. For each reaction and CA activity measurement the volumes were kept constant. Up to nine repetitions of each measurement were performed, and the time taken to decrease the pH by 0.2 units recorded. The reaction rate was then compared to the control rate, which was measured by performing the assay in the presence of an equal volume of protein and with 10% compound solvent (methanol: acetazolamide, 6-ethoxyzolamide and 1,10-phenanthroline and ethanol: diethylpyrocarbonate). This result is reported as the 100% rate, taking into considering the effect the solvents may have on CA activity. The difference between the 100% rate and that observed in the presence of the four compounds was analyzed using the Mann-Whitney U Test.

Results

Nine plasmid constructs representing nine different CA enzymes were created for this analysis (Fig. 5.5). The bacterial expression of these enzymes was achieved and confirmed by Western blot analysis and the CA activity of each determined.

5.3.1 Expression of the CA enzymes

A summary of the molecular characteristics of the CA enzymes is shown in Table 5.2. Of these 12, only nine were expressed and further characterized. Confirmation of expression was achieved by Western blotting with several antibodies, which included α-CA, α-pCA and α-His antibodies (Fig 5.7, 5.8, 5.10 – 5.15, 5.17).

Table 5.2. Characteristics of the CA enzymes analyzed.

	Length (aa)	**Lenth (aa) (+ His)**	**Predicted Mw** (kDa)**	**Observed Mw (+ His)**	**pI** (- His)**
CA1^	454	482	49.39	52	9.13
Repeat A^	762	846	27.34	28	9.42
Repeat B^*	222	247	24.01	30	7.27
Repeat C^	205	228	22.50	28	7.27
CA4^	455	484	49.35	53 and 28	9.20
CA5^	206	231	22.36	26	8.13
CA6	336	361	36.56	40	8.00
CA7	203	228	22.28	27	7.27
CA8	135	160	14.59	16	8.22
CA2	547		59.19		
CA3	655		71.33		

Mw: molecular weight; aa: amino acids; +/-His: with or without N-terminal histidine tag
^ Only these forms of the enzyme exhibited CA activity
* Repeat B had additional sequence at the C-terminus due to cloning, without this the length would be 206 aa, the Mw would be 22.46 kDa and the pI would be 8.13
**The pI and predicted molecular weight were calculated using the sequence analysis software MacVector^{TM} and did not include the N-terminal histidine tag.

5.3.1.1 CA1 expression

Bacterial expression of CA1 produced a protein with a subunit size of 52 kDa, which was similar to the predicted molecular weight (Table 5.2, Fig 5.7).

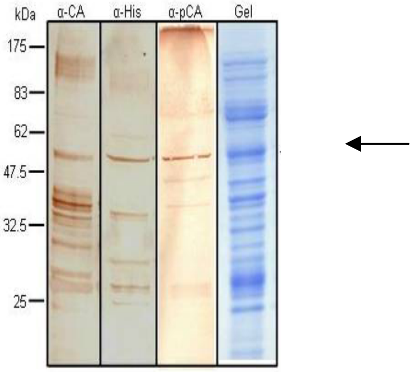


Fig. 5.7. *Western blots and Coomassie blue-stained SDS-PAGE of CA1, indicated with an arrow. 5 µl of 2.5 mg.ml^{-1} protein sample was loaded per lane (12.5 µg total protein). For each sample, 5 µl of the 200 mM imidazole eluted fraction is represented, which was then used for CA activity analysis. α-pCA refers to α-CA antibodies that have been pre-incubated with a crude extract of bacterial proteins for 3 h at 4ºC (Section 5.2.4.3).*

5.3.1.2 Repeat A expression

Repeat A was expressed with a subunit size of approximately 28 kDa as shown by Coomassie blue staining and Western blotting (Fig. 5.8). This was consistent with the predicted molecular weight based on the deduced amino acid sequence (Table 5.2).

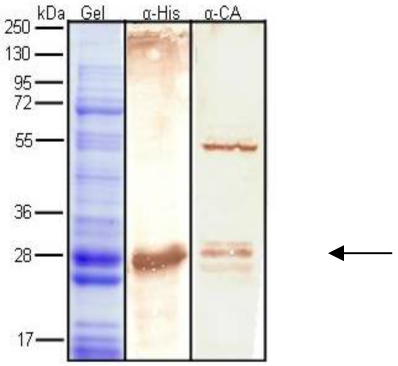


Fig. 5.8. *Coomassie blue-stained SDS-PAGE and Western blots of Repeat A, indicated with an arrow. 5 µl of 1.5 mg.ml*$^{-1}$ *protein sample was loaded per lane (7.5 µg total protein). For each sample, 5 µl of the 200 mM imidazole eluted fraction is represented, and which was then used for CA activity analysis.*

5.3.1.3 Repeat B expression

The predicted subunit size of Repeat B was 24 kDa, including the 18 exogenous amino acids present at the C-terminus (Fig. 5.9). However, when expressed the enzyme appeared to be approximately 30 kDa by SDS-PAGE and Western blotting (Fig. 5.10).

PQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMVFACSDSRVCPSVTLGLQPGEAF
TVRNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIRALLSLKDGAPDNFH
FVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEGLAGGTLKLV
GAHYDFVKGQFVTWEPP**QDALEACGTKLGCFGG***

Fig. 5.9. *The amino acid sequence of Repeat B, with the additional 16 amino acids present at the end of the sequence shown in bold. The stop codon is represented by an asterisk.*

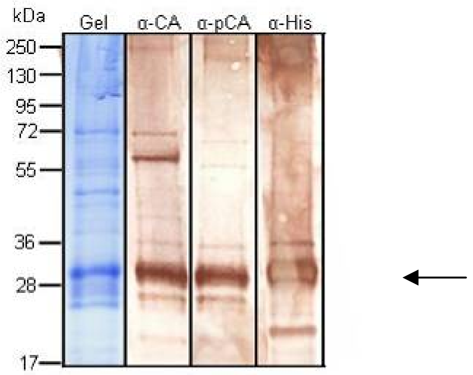


Fig. 5.10. *Coomassie blue-stained SDS-PAGE and Western blots of Repeat B, indicated with an arrow. 5 µl of 5.4 mg.ml^{-1} protein sample was loaded per lane (27 µg total protein). For each sample, 5 µl of the 200 mM imidazole eluted fraction is represented, and which was then used for CA activity analysis. α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2).*

5.3.1.4 Repeat C expression

The observed molecular weight of the Repeat C subunit was larger than predicted (Table 5.2), with a subunit size of approximately 28 kDa (Fig. 5.11).

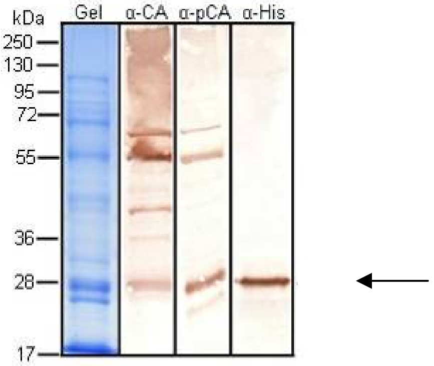


Fig. 5.11. *Coomassie blue-stained SDS-PAGE and Western blots of Repeat C, indicated with an arrow. 5 µl of protein was loaded per lane, equivalent to 10.5 µg of total protein loaded (2.1 mg.mL^{-1}). For each sample, 5 µl of the 200 mM imidazole eluted fraction is represented, and which was then used for CA activity analysis. α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2).*

5.3.1.5 CA4 expression

Expression of a construct containing cDNA corresponding to the 5′-leader sequence, Repeat A and Repeat B was predicted to produce a protein with a subunit molecular weight of 52.8 kDa (Table 5.2). However, SDS-PAGE and Western blotting showed most immuno-reactive protein at approximately 28 kDa (Fig. 5.12).

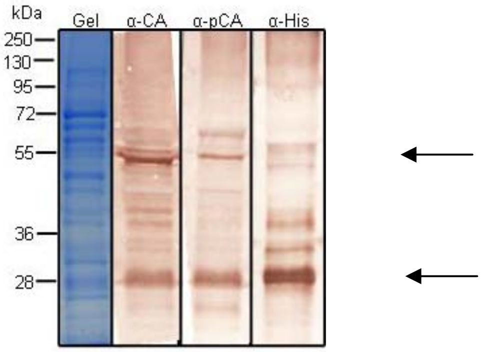


Fig. 5.12. *Coomassie blue-stained SDS-PAGE and Western blots of CA4, with the predicted molecular weight, 52.8 kDa, and the observed immuno-reactive band at 28 kDa indicated with arrows. 5 μl of the 200 mM imidazole eluted protein fraction was loaded in each lane, equivalent to 10.5 μg of total protein loaded (2.1 mg.mL*$^{-1}$*). α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2).*

5.3.1.6 CA5 expression

The CA5 construct was a chimera of Repeat A and Repeat C. The 5′-nucleotide sequence was derived from Repeat C until an *Nco*I restriction enzyme site, after which the sequence has been replaced with the 3′-end of Repeat A. The predicted molecular weight of the expressed enzyme chimera was 22.4 kDa (Table 5.2), which was similar to that observed by SDS-PAGE and Western blotting (Fig. 5.13).

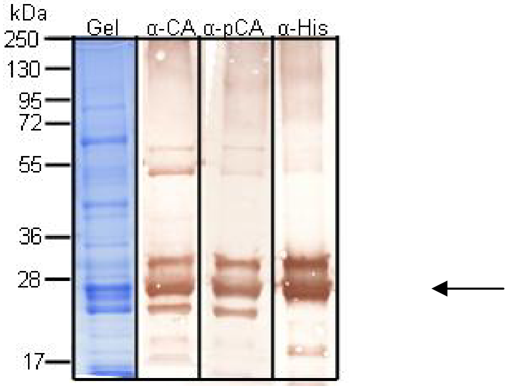


Fig. 5.13. *Coomassie blue-stained SDS-PAGE and Western blots of CA5, indicated with an arrow just below 28 kDa. 5 μl of the 200 mM imidazole eluted protein fraction was analyzed, the equivalent of 9.75 μg protein (concentration was 1.95 mg.ml^{-1}). α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2).*

5.3.1.7 CA6 expression

The CA6 construct contained cDNA sequence corresponding to the 3′-end of Repeat A (encoding 133 amino acids) adjacent to full length Repeat C. This construct was expressed with a subunit of size of approximately 40 kDa, which was close to the predicted molecular weight based on the deduced amino acid sequence (Fig. 5.14, Table 5.2). Additionally, this CA enzyme chimera displayed no CA activity.

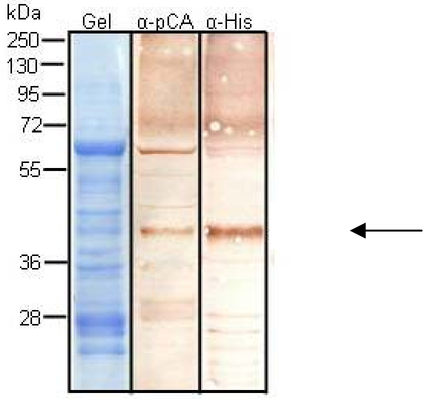


Fig. 5.14. *Coomassie blue-stained SDS-PAGE and Western blots of CA6, indicated with an arrow. 5 μl of the 200 mM imidazole eluted protein fraction was analyzed, the equivalent of 10.5 μg protein (concentration was 2.1 mg.ml^{-1}). α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2).*

5.3.1.8 CA7 expression

This enzyme was a chimeric product of Repeats A and C. It consisted of nucleotide sequence encoding 69 amino acids of the N-terminal sequence of Repeat A (but without the 5′-leader sequence) and 131 amino acids of the C-terminus of Repeat C. The predicted subunit molecular weight was approximately 22 kDa, however the observed size is slightly less than 28 kDa (Table 5.2, Fig. 5.15). The expressed enzyme displayed no CA activity.

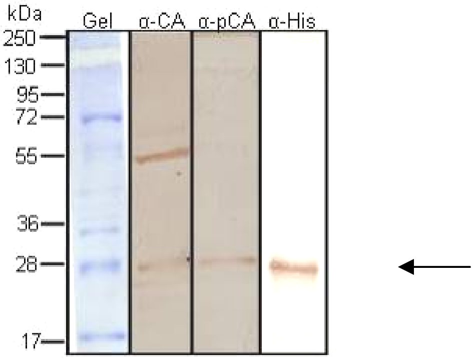


Fig. 5.15. *Coomassie blue-stained SDS-PAGE and Western blots of CA7. The protein subunit is indicated with an arrow. 5 μl of the 200 mM imidazole eluted protein fraction was analyzed, the equivalent of 12.5 μg protein (concentration was 2.5 mg.ml^{-1}). α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2).*

The effectiveness of pre-incubation of α-CA with a bacterial crude extract of proteins for reducing non-specific antibody binding was analyzed using CA7 (Fig. 5.16).

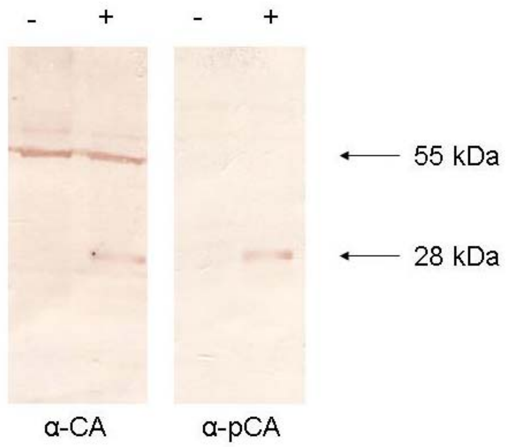


Fig. 5.16. *Western blots of the bacterial extraction of proteins without recombinant CA (-), and (+) containing CA7. Non-specific hybridization is indicated with an arrow at 55 kDa and the protein subunit is indicated with an arrow at 28 kDa. α-pCA refers to α-CA antibodies that have been pre-incubated with a crude extract of bacterial proteins for 3 h at 4ºC (Section 5.2.4.3).*

5.3.1.9 CA8 expression

This enzyme construct was a truncation of Repeat A, which contained nucleotide sequence encoding only the C-terminus of the amino acid sequence. The predicted molecular weight was 14 kDa (Table 5.2, Fig. 5.17), and this enzyme was inactive.

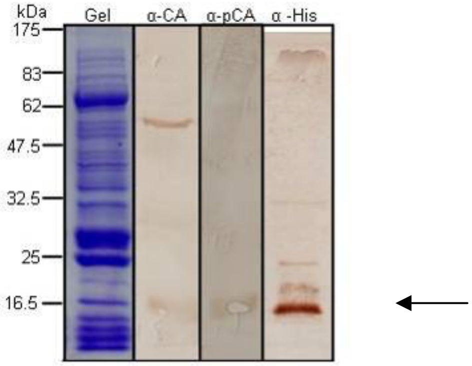


Fig. 5.17. *Coomassie blue-stained SDS-PAGE and Western blots of CA8. The protein subunit is indicated with an arrow. 5 μl of the 200 mM imidazole eluted protein fraction was analyzed, the equivalent of 15.95 μg protein (concentration was 3.2 mg.ml*$^{-1}$*). α-pCA refers to α-CA antibodies that have been blot-affinity purified (Section 5.2.4.2).*

5.3.2 Analysis of the CA enzymes

5.3.2.1 Analysis of CA activity

The activity of the expressed CA enzymes was determined as described (Section 5.2.4.5). Only five of the expressed proteins displayed CA activity (Fig. 5.18), though CA activity was detectable in a crude extract of *E. coli* expressing Repeat C. It was determined this was not due to endogenous bacterial CA activity (results not shown) although after preparation by chromatography this activity was no longer detectable. The amino acid sequence of Repeat C was compared to Repeat B (see Appendix 5.2), and the secondary structure predicted. The position of alpha-helical and beta-sheet characteristics was similar for both proteins except in one region. The sequence of Repeat B at this position (*NFHFVEDWVRIGS*) was predicted to form a beta-sheet, while the same region in Repeat C (*TFHFVEDWVKIGF*) was predicted to form an alpha-helix.

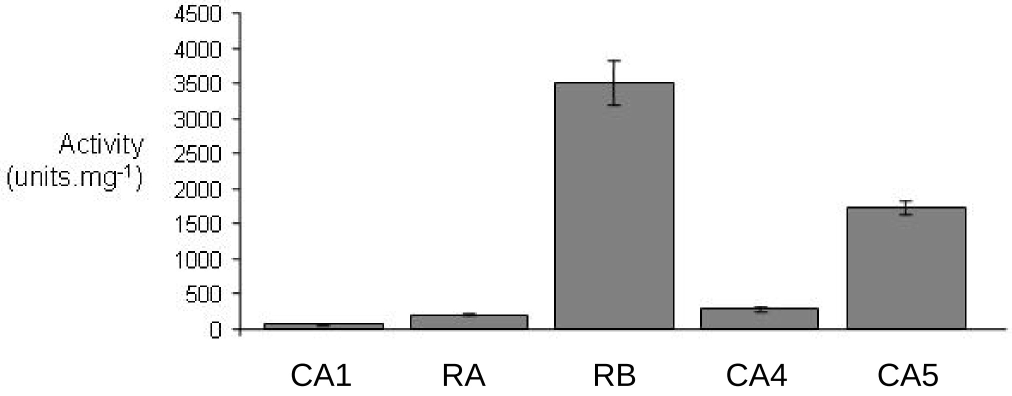


Fig. 5.18. *The activity of the bacterially expressed CA enzymes expressed as units.mg^{-1}. One unit of activity is calculated as μmol of carbon dioxide consumed per second. RA is Repeat A and RB is Repeat B.*

CA1 was assayed using 10 μl of the fraction eluted from the column with 200 mM imidazole in a 3 ml reaction volume. The protein concentration was determined by Bradford assay to be 2.5 μg.μl^{-1}, and CA1 was estimated to make up 15% of the total protein eluted by analysis of Coomassie blue-stained SDS-PAGE (Fig. 5.7). Therefore in 10 μl of protein mixture, it was determined there was 3.75 μg of CA1. The time taken to decrease the pH was recorded in triplicate, and the average and standard error determined (Fig. 5.18). The activity of CA1 was determined to be 63.3 (+/- 5.5) units.mg^{-1}.

The activity of Repeat A was determined to be 206.4 (+/-16.3) units.mg^{-1}, when 10 μl of a 1:2 dilution of the 200 mM imidazole column-eluted fraction was assayed (Fig. 5.18). From analysis of the Coomassie blue-stained SDS-PAGE, 20% of total protein (7.5 μg) was estimated to be Repeat A, resulting in 1.5 μg of protein being assayed (Fig. 5.8).

The activity of Repeat B was significantly higher than for the other expressed CA enzymes (Fig. 5.18). In order to accurately measure the protein sample, it was diluted 1:100, resulting in only 0.216 μg (where Repeat B composed 40% of the total protein sample, Fig. 5.10) being assayed. Repeat B activity was determined to be 3498.4 (+/- 313.5) units.mg^{-1}.

The activity of CA4 was determined to be 286.6 (+/- 31.8) units.mg^{-1} (Fig. 5.18). Of the 200 mM column-eluted imidazole fraction, 10 µl was assayed, corresponding to 21 µg of total protein. Analysis of CA4 by SDS-PAGE and Coomassie blue staining allowed estimation that CA4 composed 15% of the total protein sample (Fig. 5.11).

A CA chimera created to contain the N-terminus of Repeat C adjacent to the C-terminus of Repeat A (CA5) was also found to contain CA activity. CA5 was assayed using 10 µl of the 200 mM column-eluted fraction at 1:10 dilution, of which 25% was determined to be CA5 (Fig. 5.12). The activity of CA5 was therefore determined to be 1730 (+/- 100.4) units.mg^{-1} (Fig. 5.18).

5.3.2.2 Analysis of CA inhibition

The effect of four characterized CA inhibitors, acetazolamide, 6-ethoxyzolamide, 1,10-phenanthroline and diethylpyrocarbonate, on the CA activity of the expressed enzymes was determined (Table 5.3, Fig 5.19-5.23).

Table 5.3. The effect of the four compounds indicated on the enzymatic activity of the five CA enzymes expressed. The activity shown is the percent of activity remaining and is the average of the measurements taken. p was determined using a Mann-Whitney U Test.

	CA1	p	RA	p	RB	p	CA4	p	CA5	p
1mM Acetazolamide	23.7	<0.0001	37.8	0.0002	27.3	0.001	29.2	<0.0001	16.1	0.0012
1mM 6-ethoxyzolamide	23.3	<0.0001	32.6	0.0002	40.8	<0.0001	30.9	<0.0001	28.7	0.0025
10mM 1,10-phenanthroline	81.7	0.5457	91.5	0.6048	110.5	0.5457	61.8	0.0464	116.4	0.3969
1% Diethylpyrocarbonate	32.6	<0.0001	62.2	0.0019	2.7	0.0007	43.0	<0.0001	22.5	<0.0001

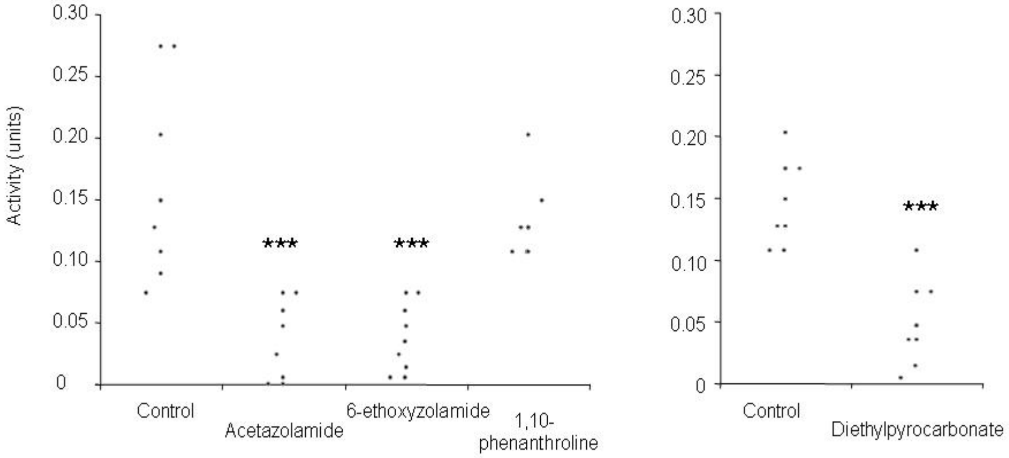


Fig. 5.19. *The activity of CA1 (expressed as units) in the presence of 1 mM acetazolamide, 1 mM 6-ethoxyzolamide, 10 mM 1,10-phenanthroline or 1% diethylpyrocarbonate. The measurements taken are compared to the control reactions, which were performed in the presence of 10% methanol (acetazolamide, 6-ethoxyzolamide and 1,10-phenanthroline) or 10% ethanol (diethylpyrocarbonate). *p= <0.05, ** p=<0.01, *** p=<0.001 Mann-Whitney U Test.*

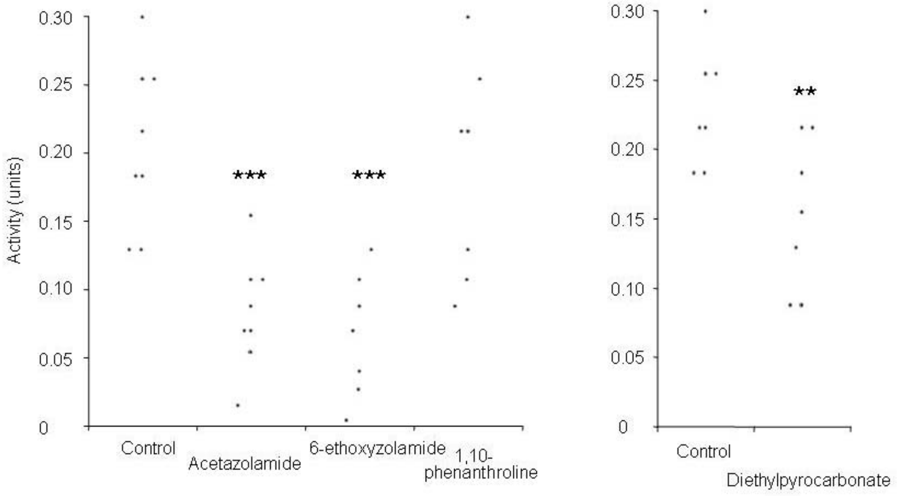


Fig. 5.20. *The activity of Repeat A (expressed as units) in the presence of 1 mM acetazolamide, 1 mM 6-ethoxyzolamide, 10 mM 1,10-phenanthroline or 1% diethylpyrocarbonate. The measurements taken are compared to the control reactions, which were performed in the presence of 10% methanol (acetazolamide, 6-ethoxyzolamide and 1,10-phenanthroline) or 10% ethanol (diethylpyrocarbonate). *p= <0.05, ** p=<0.01, *** p=<0.001 Mann-Whitney U Test.*

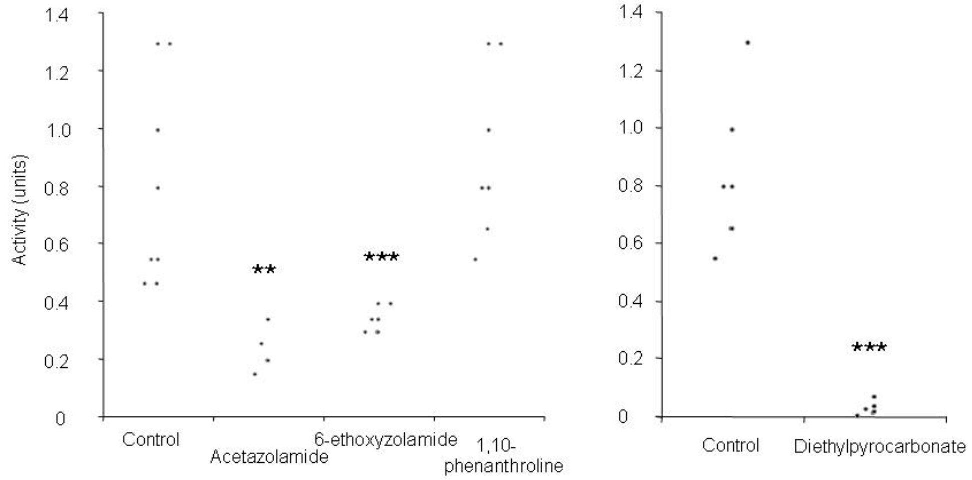


Fig. 5.21. *The activity of Repeat B (expressed as units) in the presence of 1 mM acetazolamide, 1 mM 6-ethoxyzolamide, 10 mM 1,10-phenanthroline or 1% diethylpyrocarbonate. The measurements taken are compared to the control reactions, which were performed in the presence of 10% methanol (acetazolamide, 6-ethoxyzolamide and 1,10-phenanthroline) or 10% ethanol (diethylpyrocarbonate). *p= <0.05, ** p=<0.01, *** p=<0.001 Mann-Whitney U Test.*

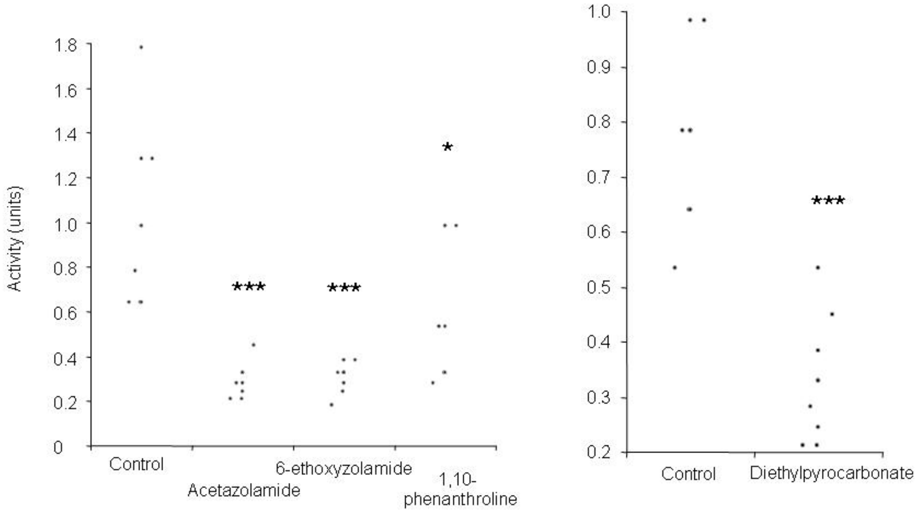


Fig. 5.22. *The activity of CA4 (expressed as units) in the presence of 1 mM acetazolamide, 1 mM 6-ethoxyzolamide, 10 mM 1,10-phenanthroline or 1% diethylpyrocarbonate. The measurements taken are compared to the control reactions, which were performed in the presence of 10% methanol (acetazolamide, 6-ethoxyzolamide and 1,10-phenanthroline) or 10% ethanol (diethylpyrocarbonate). *p= <0.05, ** p=<0.01, *** p=<0.001 Mann-Whitney U Test.*

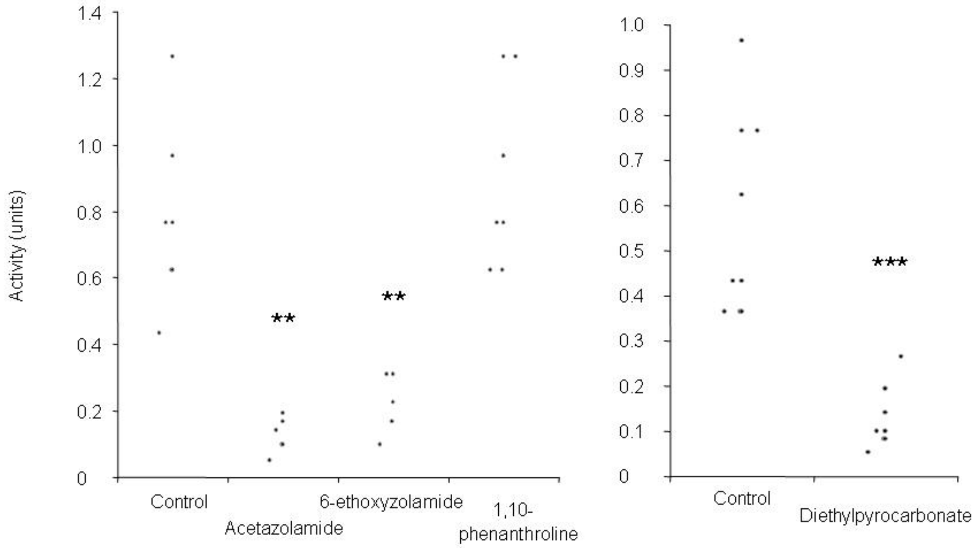


Fig. 5.23. *The activity of CA5 (expressed as units) in the presence of 1 mM acetazolamide, 1 mM 6-ethoxyzolamide, 10 mM 1,10-phenanthroline or 1% diethylpyrocarbonate. The measurements taken are compared to the control reactions, which were performed in the presence of 10% methanol (acetazolamide, 6-ethoxyzolamide and 1,10-phenanthroline) or 10% ethanol (diethylpyrocarbonate). *p= <0.05, ** p=<0.01, *** p=<0.001 Mann-Whitney U Test.*

Discussion

Nine plasmid constructs representing nine different CA enzymes were created and expressed in a bacterial expression system, and the CA activity determined for each. Expression was confirmed by SDS-PAGE and Western blotting. Of those that displayed CA activity, the response to common CA inhibitors was also analyzed. Each Repeat (Repeat A, Repeat B and Repeat C) was expressed in order to determine whether these could independently catalyze the hydration of carbon dioxide, as well as several combinations of the components of the CA transcript to determine what influence these parts of the transcript had on CA activity. The 276 bp insert unique to CA2, or the full length CA2 or CA3 transcripts were not expressed as recombinant proteins. When amplified from maize leaf RNA or cDNA, the 276 bp insert of CA2 contained nucleotide sequence differences in the open reading frame that would attenuate translation (Section 6.3.3, Chapter 6; Genbank accession U08401, version gi:606810).

5.4.1 CA expression

Under the same conditions for bacterial cell growth and protein expression, CA1 was expressed at relatively low levels compared to the other enzymes, particularly Repeat B (Fig. 5.7, Fig. 5.10). This may be due to degradation of the translated peptide, as when analyzed using α-His antibodies there are several immuno-reactive bands that may represent truncated peptides still containing the N-terminal histidine tag (Fig. 5.7). However, a distinct band representing the recombinant protein at 52 kDa was distinguishable with Coomassie blue staining and immunoblotting. Repeat A was expressed with a subunit size of approximately 28 kDa as shown by Coomassie blue staining and Western blotting (Fig. 5.8). This was consistent with the predicted molecular weight based on the deduced amino acid sequence (Table 5.2). It would appear that in the maize leaf, expression of Repeat A including the amino acid sequence encoded by the 5′-leader sequence could be the source of the 28 kDa band identified by Western blotting of crude maize leaf extract (Section 6.3.4.1, Chapter 6; Burnell and Ludwig, 1997). An immuno-reactive band is observed at 55 kDa when the blot is probed with α-CA, however this is most likely non-specific antibody reactivity.

Repeat B was expressed at the highest levels under the same conditions as used for the other CA enzymes (Fig. 5.10). Both Repeat B and Repeat C appeared to be larger than expected based on the predicted molecular weight using the deduced amino acid sequences. The subunit size of Repeat B was predicted to be 24 kDa, while Repeat C was predicted to be 22.5 kDa; however Repeat B appeared to be approximately 30 kDa while Repeat C was closer to 28 kDa (Fig. 5.10, Fig. 5.11, Table 5.2). The discrepancies in the predicted and observed molecular weights are most likely due to the inherent inaccuracies in determining subunit size by SDS-PAGE. Additionally when rice CA, which is very similar to maize CA (Fig. 5.24), was expressed using a bacterial expression system as a glutathione *S*-transferase fusion protein, it migrated at the predicted molecular weight indicating that no modification of the translated peptide was occurring (Yu *et al.,* 2007).

```
maize           MDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVCPSVTLGLQPGE 60
rice            MDAAVDRLKDGFAKFKTEFYDKKPELFEPLKAGQAPKYMVFSCADSRVCPSVTMGLEPGE 60
                **.:*:***.** *****.************:**:*:****:*:*********:**:***

maize           AFTVRNIASMVPPYDKIKYAGTGSAIEYAVCALKVQVIVVIGHSCCGGIRALLSLKDGAP 120
rice            AFTVRNIANMVPAYCKIKHAGVGSAIEYAVCALKVELIVVIGHSRCGGIKALLSLKDGAP 120
                ********.***.* ***:**.*************::******* ****:**********

maize           DNFTFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEGLAG 180
rice            DSFHFVEDWVRTGFPAKKKVQTEHASLPFDDQCAILEKEAVNQSLENLKTYPFVKEGIAN 180
                *.* ******* * ***:**:.****:******:******** **:***:*******:*.

maize           GTLKLVGAHSHFVKGQFVTWEP 202
rice            GTLKLVGGHYDFVSGNLDLWEP 202
                *******.* .**.*::  ***
```

Fig 5.24. *ClustalW alignment of the deduced amino acid sequences of Repeat A of the maize CA2 cDNA sequence and the rice chloroplastic CA (not including the chloroplast transit peptide).*

Expression of a construct containing amino acid sequence encoded by the 5′-leader sequence, Repeat A and Repeat B (CA4) was predicted to produce a protein with a subunit molecular weight of 49.3 kDa (Table 5.2). However, SDS-PAGE and Western blotting showed most immuno-reactive protein at approximately 28 kDa (Fig. 5.12). Like CA1, the translated peptide may be subjected to degradation in the bacterial expression system. However, the processed subunit size is the same as the subunit size of Repeat A at 28 kDa (Table 5.2). There are no obvious processing sites between Repeat A and Repeat B in the amino acid sequence and it may also be that the translated peptide was cleaved in half, particularly as the measured CA activity was not equivalent to the sum of that observed for Repeat A and Repeat B when expressed separately (Fig. 5.18).

The CA5 construct was a chimera of Repeat C and Repeat A, which effectively resulted in expression of Repeat B, however without the additional exogenous amino acids at the C-terminus of the protein (Fig. 5.5, Fig. 5.13). CA6 contained the C-terminus of Repeat A (133 amino acids) adjacent to full length Repeat C. This chimera was expressed with a subunit of size of 40 kDa, which was only slightly larger than the predicted molecular weight based on the deduced amino acid sequence (Fig. 5.14, Table 5.2). Additionally, CA6 displayed no CA activity. CA7, which was a chimeric product of Repeats A and C also displayed no CA activity. CA7 consisted of 69 amino acids of the N-terminal sequence of Repeat A (without amino acid sequence encoded by the 5′-leader sequence) and 131 amino acids of the C-terminus of Repeat C. The predicted subunit molecular weight was approximately 22 kDa, however the observed size is slightly less than 28 kDa (Fig. 5.15, Table 5.2). Expression of truncated Repeat A (CA8) produced a protein with a subunit molecular weight of approximately 14 kDa (Fig. 5.17), and this protein contained no CA activity.

Rice CA has been expressed in a recombinant system and the expressed protein characterized (Yu *et al.,* 2007). From the deduced amino acid sequence, it was determined that this CA was basic, with a pI value of 8.41. The maize CAs also have a basic calculated pI value, and inclusion of the deduced amino acid sequence of the 5′-leader sequence in the calculation resulted in a pI value of greater than 9 (Table 5.2). This is in contrast to the two CA isoforms purified from pea, which both had pI values below 7, at 5.75 and 6.3.

5.4.2 Immuno-reactivity of the CA enzymes

The protein-specific antibodies used in this analysis were generated using bacterially expressed and purified CA1 (Burnell Laboratory, James Cook University, Australia). Despite being generated against a purified peptide, it appeared that the polyclonal antibody serum also contained antibodies against many bacterial proteins that would have been associated with CA1 after expression in a bacterial system. In most cases a 55 kDa protein present in the bacterial extract was highly reactive with α-CA (Fig. 5.7, 5.8, 5.10 – 5.15, 5.17). It was thought this could represent an *E. coli* beta-CA, however the *E. coli* genome has been sequenced and the two beta-CAs characterized (*cynT* and *cynT2*) are only 219 and 220 amino acids respectively and have predicted molecular weights of approximately 24 kDa (www.ncbi.nlm.nih.gov, accession: BAE76121 and P61517, Cronk *et al.,* 2001).

In order to reduce the non-specific antibody reactivity, the antibodies were enriched by blot-affinity purification as well as by pre-incubation with a crude bacterial extract. Blot-affinity purification involved purifying only those antibodies that were reactive against the protein of interest (Burnell and Ludwig, 1997). This was done by allowing the antibodies to bind Repeat B, then eluting them from the membrane using glycine buffer at a very low pH (Section 5.2.4.2). Incubating α-CA with a crude bacterial extract of proteins was intended to remove non-specific antibodies before the antibody was used (Section 5.2.4.3). For CA1 and CA7, pre-incubating α-CA with a bacterial crude extract was very effective for eliminating non-specific antibody binding (Fig 5.7, Fig. 5.16). For Repeat B, Repeat C, CA4 and CA5 blot-affinity purification of the antibodies only slightly reduced non-specific reactivity (Fig. 5.10 – 5.13).

5.4.3 Analysis of CA activity

The CA activity of each of the expressed enzymes was measured to determine if there were any differences in enzymatic activity despite the similarities in primary structure. The activity of Repeat A was significantly higher than CA1 and was determined to be 200 units.mg^{-1} (Fig. 5.18). Repeat B displayed the highest CA activity at 3,395 units.mg^{-1}, however this may be an overestimate based on the method used to quantify the protein assayed (Section 5.2.4.4). Unlike Repeat B, Repeat C had very low CA activity, which was only detectable in the crude bacterial extract and after concentrating the protein by affinity chromatography this activity could no longer be detected. An alignment of the active and inactive CA sequences was made to determine what differences were evident that may account for the decreased or lack of activity (Fig. 5.25). The amino acid sequences of three active CAs were aligned with two inactive CAs, and key differences in residues at the C-terminus of the protein were identified.

```
                                                    o        ● *
RB              MDPPQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMVFACSDSRVCPSVTLGLQ
CA5             MDPPQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMVFACSDSRVCPSVTLGLQ
RA              ---MDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVCPSVTLGLQ
RC              MDPPQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMVFACSDSRVCPSVTLGLQ
CA7             ---MDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVCPSVTLGLQ

                                             *               ●  ●
RB              PGEAFTVRNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIRALLSLKD
CA5             PGEAFTVRNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIRALLSLKD
RA              PGEAFTVRNIASMVPPYDKIKYAGTGSAIEYAVCALKVQVIVVIGHSCCGGIRALLSLKD
RC              PGEAFTVRNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIRALLSLQD
CA7             PGEAFTVRNIASMVPGYDKTKYTGIGSAIEYAACALKVEVLVVIGHSCCGGIRALLSLQD

                    ▬▬▬▬▬▬▬▬▬▬▬▬▬                  △      *
RB              GAPDNFHFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEG
CA5             GAPDNFHFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEG
RA              GAPDNFHFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEG
RC              GAPDTFHFVEDWVKIGFIAKMKVKKEHASVPFDDQCSILEKEAVNVSLENLKTYPFVKEG
CA7             GAPDTFHFVEDWVKIGFIATMKVKKEHASVPFDDQCSILEKEAVNVSLENLKTYPFVKEG

RB              LAGGTLKLVGAHYDFVKGQFVTWEPPQDALEACGTKLGCFGG
CA5             LAGGTLKLVGAHYDFVKGQFVTWEP-----------------
RA              LAGGTLKLVGAHYDFVKGQFVTWEPP----------------
RC              LANGTLKLIGAHYDFVSGEFLTWKK-----------------
CA7             LANGTLKLIGGHYDFVSGEFLTWKK-----------------
```

Fig. 5.25. *Alignment of the active, expressed CA amino acid sequences (RB – Repeat B, CA5 – Repeat C/A chimera, and RA – Repeat A) with the inactive amino acid sequences (RC – Repeat C and CA7 – Repeat A/C chimera). Differences in amino acid residues for both groups of sequences are indicated with bolding. Asterisks indicate potential zinc ligands, and filled circles indicate potential zinc ligands that when mutated, resulted in CA that bound zinc poorly (Bracey* et al.*, 1994; Cronk* et al.*, 2001). The open circle indicates the glutamine residue suggested to act as a hydrogen bond donor in the Arabidopsis beta-CA (Rowlett* et al.*, 1994). The triangle indicates a cysteine residue that is conserved between monocot and dicot plant sequences, which in pea is essential for structure and enzyme activity (Björkbacka* et al.*, 1997). The predicted secondary structure was different for the active and inactive sequences in the region indicated by the grey line.*

The differences in the amino acid sequences between active and inactive CAs are predominantly at the C-terminus of the peptide sequence (Fig. 5.25). At the N-terminus, there are no differences that are unique to only the active or inactive CAs. Despite containing all the amino acids that have been identified as being putative zinc ligands, or those residues that have been shown to be important for enzyme structure and activity (Bracey *et al.,* 1994;

Cronk *et al.,* 2001; Rowlett *et al.,* 1994; Björkbacka *et al.,* 1997), Repeat C (RC) and the Repeat A/C chimera (CA7) did not show detectable or measurable CA activity. The difference between the properties of the amino acids at the C-terminus was therefore compared.

The most significant difference appears to be that the C-terminus of the inactive CAs contained many non-polar and therefore hydrophobic amino acid residues, such as phenylalanine, isoleucine, methionine and leucine. Due to the hydrophobic nature of these amino acids, they may change the secondary structure of the expressed peptides significantly enough to cause reduced CA activity. In support of this, when the predicted secondary structure of Repeat C was compared with Repeat B, a region of difference was identified. Repeat C had alpha-helical characteristics at this position, and Repeat B contained a beta-sheet (Fig. 5.25). The active CAs also contain more small amino acids such as asparagine, serine, proline, glycine, and valine. Importantly, the final two amino acids in the active CA sequences are glutamate, while in the inactive CAs the presence of two lysine residues creates a positive charge at the C-terminus. It may be that the C-terminus of the protein is necessary for monomer association enabling formation of the catalytic site, which for beta-CAs is typically at the interface of two monomers. The crystal structure of the CA from pea shows that the monomer has two protruding motifs, corresponding to the N-terminus and C-terminus. These mediate the interactions necessary for oligomerization, specifically association as an octamer, rather than for the formation of dimers (Kimber and Pai, 2000).

When Repeat A and Repeat B were expressed separately, measurable CA activity was detected (Fig. 5.18), and analysis of amino acid sequence confirmed that Repeat A and Repeat B contain the putative catalytic amino acids. With the identification of two immuno-reactive CA subunits in maize leaf tissue that are 27 kDa and 28 kDa (Burnell and Ludwig, 1997), it is highly likely that in the plant, Repeat A and Repeat B exist independently as active CA isozymes. However, the transcripts encoding these isozymes also contain Repeat C included in the open reading frame. Repeat C appears to have little CA activity, and therefore the purpose of the Repeat C peptide is unclear.

It may be that the low level of activity observed for Repeat C was due to the method used to measure enzymatic activity. An *E. coli* CA, as well as a beta-CA from *Mycobacterium tuberculosis* were inactive when assayed at neutral pH, though contained significant activity above pH 8 (Covarrubias *et al.,* 2005; Cronk *et al.,* 2001). In *Amaranth cruentus* leaves, two CAs exist that display highest V_{max} values at pH 7.6. One of these has low affinity to carbon dioxide at this pH, with the K_m decreasing up to pH 9 (Guliev *et al.,* 2003). Alternatively, the association of Repeat C to either Repeat A or Repeat B may be a regulatory mechanism that renders the enzyme inactive, and cleavage of Repeat C allows the isozymes to form active enzyme complexes. This would be a novel regulatory mechanism for this enzyme, and no other examples of this are reported in the literature. Rather the CA enzymes that have been found to contain two protein domains on a single peptide, including those from either the alpha- or beta-CA gene families, are assumed to form an active enzyme in a similar arrangement that would arise from the association of two monomers (Mitsuhashi *et al.,* 2000; Covarrubias *et al.,* 2005).

When purified from leaf tissue, a CA isozyme that had a native molecular weight of approximately 180 kDa was found to dissociate into 47 kDa subunits, which were detectable by measuring CA activity (Burnell and Ludwig, 1997). The components of the CA transcript that are translated to form this 47 kDa subunit remain unknown. From the molecular data available it can be presumed that a 44 kDa subunit would be generated by expression of Repeat A and Repeat C or Repeat A and Repeat B as a single peptide without translation of the 5′-leader sequence, or a 49 kDa subunit including the amino acid sequence encoded by the 5′-leader sequence (Table 5.2). CA1 and CA4, both of which included the amino acid sequence encoded by the 5′-leader sequence, were expressed and found to contain relatively low levels of CA activity (Fig. 5.18). While the activity of the 47 kDa isoform purified from leaf tissue must have been significant in order to be detectable (Burnell and Ludwig, 1997), both of the bacterially expressed peptides were particularly susceptible to proteases, and therefore the enzymatic activity of these enzymes was most likely not accurately depicted (Fig. 5.7, Fig. 5.12, Fig. 5.18). This is in agreement with that observed for the 52 kDa CA subunit identified in Western blots of maize leaf protein, which was nearly undetectable in the absence of protease inhibitors (Burnell and Ludwig, 1997). It appears that the CA peptide containing more than one Repeat, and therefore more than one protein domain, is particularly unstable.

5.4.4 Analysis of CA inhibition

CA isozymes from the three main gene families, alpha, beta and gamma, are commonly inhibited by two groups of compounds, metal anions and sulphonamides, with much research focusing on the inhibition of the mammalian alpha-CAs due to their involvement in physiological and pathological processes. For example, acetazolamide is a sulphonamide that is used as a pharmaceutical ('Diamox') for the treatment of a range of conditions from glaucoma to altitude sickness (Supuran *et al.*, 2003). CAs from the alpha-gene family have been well characterized with crystal structures for many of the enzymes resolved and the reaction mechanism clearly defined (Fig. 5.4). The zinc ion present in all CAs is essential for catalysis, and both anions and sulphonamides inhibit CA by forming a complex with the catalytic zinc (Vidgren *et al.,* 1990).

The sensitivity of the expressed active CA enzymes to acetazolamide as well as 6-ethoxyzolamide was compared (Table 5.3, Fig 5.19-5.23). All five CAs show a loss of approximately 70% activity in the presence of these compounds. It is possible that the CA enzymes that contained two protein domains on a single peptide, CA1 and CA4, were inhibited slightly more than the single domain enzymes, Repeat A, Repeat B and CA5, although the difference is not significant. The crystal structure of the pea CA has shown that the catalytic zinc ion exists in a hydrophobic pocket and is not easily accessible (Kimber and Pai, 2000). If expression as a dual domain peptide disturbs the secondary structure of the enzyme, considering that the activity of CA1 and CA4 was equal to neither the sum nor the combined activity of the individual peptide components, Repeat A, Repeat B or Repeat C, then it may be possible that these inhibitors could gain greater access to the active site.

Inhibition of the active CA enzymes was also investigated in the presence of 1,10-phenanthroline, which chelates heavy metal ions and has been used to characterize CA isolated from *A. cruentus.* A pH- and temperature-dependent inhibitory effect was observed for the *A. cruentus* CA isoforms, with highest inhibition demonstrated at low or high pH levels, while the activity was close to control rates when measured at neutral pH (Guliev *et al.,* 2003). The activity of the single domain peptides, particularly Repeat B and CA5, in the presence of 1,10-phenanthroline was greater than that observed for the control reactions (Table 5.3). As 1,10-phenanthroline chelates heavy metal ions, and these have been shown to be inhibitory to CA activity, it may be that the compound was chelating the inhibitory anions present in the buffers and solutions used to assay CA activity. This effect is not seen for CA1 and CA4 again suggesting a different secondary structure, which may allow access to the active site resulting in chelating of the catalytic zinc and a resulting loss of CA activity (Fig. 5.19 and Fig. 5.22).

Inhibition of CA in the presence of anions may also have affected the CA activity measurements described in this analysis (Fig. 5.18; Atkins *et al.,* 1972b; Ivanov *et al.,* 2007). For example, chloride ions were present in the both the protein buffer and in the buffer used to measure activity. Distinct from human CAII, the beta-CA from pea was significantly bound by chloride as well as sulfate ions, which had an inhibitory effect particularly at low pH (Johansson and Forsman, 1993). It has been suggested that these anions would also be present in a typical cell environment, and that the maximum rate of CA activity observed *in vivo* may be constrained by competitive inhibition, rather than being a reflection of the maximum turnover rate (Kimber and Pai, 2000).

The proton transfer step of the CA reaction mechanism has been described as the rate-limiting step (Johansson and Forsman, 1993; Rowlett *et al.,* 1994). For alpha-CA enzymes both protein crystallography and mutagenesis studies have identified a histidine residue (His64 in human CAII) as the proton acceptor; however whether this is also the case for beta-CAs is unknown. Two potential amino acids have been identified in the Arabidopsis beta-CA, His216 and Tyr212, both of which when mutated significantly effected the proton transfer step of the reaction (Rowlett *et al.,* 2002). This histidine residue is eleven amino acids N-terminus of the putative zinc-binding histidine residue, and in the maize CA sequence corresponds to an alanine residue. However, the tyrosine residue at position 212 in Arabidopsis is also conserved in the maize sequence (Fig. 5.25).

The role of the histidine residues in CA was investigated by measuring enzyme activity in the presence of diethylpyrocarbonate, which modifies histidine. For beta-CAs the catalytic zinc ion is coordinated by a histidine residue as well as two cysteine residues (Kimber and Pai, 2000). The most significant inhibition is observed for Repeat B (Fig. 5.21), which may indicate an inability for the enzyme to bind zinc in the presence of diethylpyrocarbonate. However the other enzymes do not appear to be as severely effected

(Table 5.3). This raises the question of whether Repeat B has a different reaction mechanism or secondary structure to Repeat A or Repeat C. That CA4 is not inhibited like Repeat B in the presence of diethylpyrocarbonate, but rather like Repeat A, implied that the activity observed for this enzyme came from the Repeat A protein domain, rather than Repeat B, and this was similarly observed in the presence of 1,10-phenanthroline (Table 5.3).

Conclusion

CA from maize and other C_4 monocot species are distinct from other beta-CAs as they are encoded by large transcripts that in maize are due to the presence of repeating sequences that represent multiple protein domains (Burnell and Ludwig, 1997). Expression of nine CAs in a bacterial expression system allowed for characterization of the molecular weight, enzymatic activity and inhibitor sensitivity of the enzyme.

The expression of nine different CAs was confirmed by SDS-PAGE and analysis by Western blotting with antibodies raised against bacterially expressed CA1. Enzymatic activity of CA1, which is encoded by the 5′-leader sequence, Repeat A and Repeat C was confirmed, and significantly the individual protein domains, particularly Repeat A and Repeat B, display independent CA activity. Repeat C, however, appeared to be relatively inactive compared to Repeat A and Repeat B, and the sequence differences between the active and inactive forms was analyzed. It was found that Repeat C contained differences in the amino acid sequence that resulted in a change in the predicted secondary structure, and possibly a different oligomerization state, as evidenced by the presence of two positively charged amino acids at the C-terminus of the protein.

In future work, the identification of exactly which amino acids are responsible for the reduced activity observed for Repeat C could be determined by creating more chimeras, focusing on the amino acids identified from sequence alignments. Also, experiments to determine the pH-dependence of V_{max} and K_m values would be informative for further characterizing the maize enzyme. Ultimately, plant-purified CA protein sequence, for example obtained by mass spectrometry, is needed to clarify which components of the transcript are being expressed in the maize leaf. Several attempts were made to purify CA from maize leaf tissue, but these were unsuccessful (Section 6.3.4, Chapter 6).

Finally, measuring the inhibition of CA in the presence of the sulphonamides, acetazolamide and 6-ethoxyzolamide, as well as in the presence of 1,10-phenanthroline and diethylpyrocarbonate provided some insights into the reaction mechanism of maize CA, as well as the structure of the enzyme when expressed as a single peptide with two protein domains. Generating CA with mutations at amino acids responsible for binding the catalytic zinc ion as well as those identified as possibly being involved in the proton transfer step of the reaction mechanism would enable a comparison of the C_4 monocot beta-CAs with CAs from C_3 dicot species, to determine if a different reaction mechanism is employed.

CHAPTER 6

6. Analysis of the Maize CA Isozymes

Introduction

Maize as well as sorghum and sugarcane are grasses and are classified as C_4 monocots, developing with only one seed leaf. To date, the CAs from these agronomically important species remains relatively uncharacterized. This is in contrast to CA from dicot species that use the C_3 photosynthetic pathway where CA sub-cellular localization, protein structure and physiological function have been resolved.

6.1.1 CA in monocot and dicot plant species

Differences between monocot and dicot CAs were first identified based on the patterns observed after polyacrylamide gel electrophoresis (Atkins, *et al.* 1972a). A comparison of monocot and dicot CA amino acid sequences revealed three differences:

- The transit peptide of monocot CAs is approximately forty amino acids shorter than for dicot CAs;
- Dicot sequences have a ten amino acid extension at the C-terminal end of the protein; and
- 12 amino acid residues are homologous across monocot species that are different from and that are conserved across dicot species (Burnell, 2000).

The differences between monocot and dicot beta-CAs also include differences in molecular weight, quaternary structure, and sensitivity to inhibitors and reducing environment. Monocot CAs were first reported as having a molecular weight of approximately 40 kDa and were predominantly monomers, which was similar to the animal CAs (Graham *et al.,* 1984). Two 180 kDa isoforms were later identified in the C_4 monocot maize, which had subunit sizes of 28 and 47 kDa, while the CA from the C_3 monocot rice had a molecular weight of 29 kDa (Burnell and Ludwig, 1997; Yu *et al.,* 2007). The dicot CAs were thought to have native molecular weights of approximately 180 kDa, with the association of several subunits required to form active enzyme (Atkins *et al.,* 1972a). It has been suggested that the ten amino acid extension at the C-terminus of the dicot CA sequence may be involved in oligomerization (Burnell, 2000). The quaternary structure of CAs from C_3 dicot species has since been reported to be either a hexamer or octamer, with molecular masses ranging from 140 kDa to 250 kDa (Reed and Graham, 1980; Johansson and Forsman, 1993). Also, the crystal structure of the pea CA shows two protruding motifs, corresponding to the N-terminus and C-terminus of the monomer, and these mediate the interactions that are necessary for oligomerization (Kimber and Pai, 2000).

When purified from pea leaf tissue (a C_3 dicot), CA activity was affected by reducing agents and the enzyme appeared to be an octamer (Johansson and Forsman, 1993). This was confirmed by structural X-ray analysis (Kimber and Pai, 2000). The association of enzyme subunits to form the native quaternary state was dependent on two cysteine residues at positions 269 and 272 in the pea CA amino acid sequence. The first of these is conserved in both monocot and dicot beta-CA sequences, while the second is only present in dicot CAs and may explain why monocot CAs are not as sensitive to oxidation (Björkbacka *et al.,* 1997). The identification of conserved cysteine residues is often an indication that the protein is subject to redox regulation, and a partially purified chloroplastic beta-CA showed at least a two-fold increase in activity in the presence of glutaredoxin (Rouhier *et al.,* 2005).

In addition, immunological experiments have indicated that there are antigenic differences between CAs from monocot and dicot species. Mono-specific antibodies raised against CA from the dicot spinach were more reactive against CA from dicot plant extracts than from monocot plant extracts (Okabe *et al.,* 1984). In contrast, antibodies raised against monocot CA from maize were cross-reactive with leaf extracts from a variety of other dicot species, but only quantitatively titrated CA activity from monocot plant extracts (Burnell, 1990).

6.1.2 Localization of CA in C_3 and C_4 plants

CA is located in the chloroplasts of mesophyll cells in the leaves of C_3 plants, such as rice, where it accounts for up to 2% of total leaf protein (Okabe *et al.,* 1984). However, in plants using the C_4 photosynthetic pathway, CA activity in the chloroplasts, particularly in the bundle sheath cells surrounding the vascular tissue, would be detrimental to the operation of the C_4 pathway (Poincelot, 1972; Burnell and Hatch, 1988). CA activity in the alkaline environment of the chloroplast would result in the conversion of carbon dioxide to bicarbonate, effectively reducing the supply of carbon dioxide for Rubisco. The C_4 pathway is usually a dual-celled system allowing for separation of the processes that occur in the external mesophyll cells and the inner bundle sheath cells, which surround the vascular tissues (Edwards *et al.,* 1985). Any CA activity that has been detected in chloroplast or bundle sheath cell extracts has been assumed to be the result of sample contamination from the mesophyll cell cytoplasm, where CA is predominantly located in C_4 plants (Burnell, 2000).

Analysis of the C_4 dicot species *Flaveria bidentis* CA cDNA sequences showed the presence of a chloroplast transit peptide that was not processed after translation, confirming the localization of CA in C_4 plants to the mesophyll cell cytosol (Cavallaro *et al.,* 1994). Three CA isozymes exist in this species, and one was imported into chloroplasts in *in vivo* chloroplast uptake assays (Tetu *et al.,* 2007). Similarly, two CA isozymes were purified from the leaves of *Amaranthus cruentus,* a C_4 plant, one of which was associated with the chloroplast membrane of the bundle sheath cells (Guliev *et al.,* 2003). Here, it was proposed that CA played a role in facilitating transport of carbon dioxide across the membrane to be in the vicinity of the carboxylation enzymes. In Arabidopsis, six CA isoforms exist that are encoded by different genes and exist in different sub-cellular compartments (Fabre *et al.,* 2007). Using antibodies against the pea CA, a hybridizing band was detected from a cytosolic extract, indicating that in C_3 plants CA also exists in the cell cytoplasm (Fett and Coleman, 1994).

Thylakoid membranes from several plant species display reversible carbon dioxide hydration activity associated with photosystem II and the electron transport chain (Stemler, 1997). In maize, a chloroplastic 33 kDa protein was immuno-reactive with antibodies generated against a thylakoid alpha-CA from *Chlamydomonas reinhardtii* and in pea, a 25 kDa protein was associated with Rubisco on the outer surface of the thylakoid membranes (Lu and Stemler, 2002; Lazova and Stemler, 2008). These findings suggest a role for alpha-CA in the chloroplasts, rather than beta-CA.

6.1.3 Structure of plant CA

One of the earliest CA enzymes was purified from the leaves of pea and appeared to be a hexamer containing six zinc atoms (Kisiel and Graf, 1972; Reed and Graham, 1981). The deduced molecular mass of two mature polypeptides was 24 kDa and 28 kDa (Roeske and Ogren, 1990; Majeau and Coleman, 1992). When analyzed by circular dichroism (CD) spectroscopy, pea CA had predominantly alpha-helical characteristics (Mitsuhashi *et al.,* 2000). In spinach, CA is a hexamer with a zinc ion per enzyme subunit. Each subunit has a deduced molecular weight of approximately 35 kDa, which includes a chloroplast transit peptide (Pocker and Ng, 1973; Fawcett *et al.,* 1990). CA was purified from spinach leaf tissue and shown to have a molecular weight of 26.2 kDa, with a 3.5 kDa transit peptide directing localization to the chloroplast (Burnell *et al.,* 1990a). The first CA identified in *F. bidentis* also contained a chloroplast transit peptide, and the deduced molecular weight including the transit peptide was 35 kDa (Cavallaro *et al.,* 1994). The CA purified from *Cicer arietinum* (chick pea) leaf tissue was an octamer, containing one zinc atom per 26 kDa subunit (Guliev *et al.,* 2003).

The crystal structures of several beta-CAs have been determined, and these are predominantly composed of alpha-helices (Fig. 6.1 – 6.3). The pea CA is made up of a novel arrangement of dimers of dimers of dimers, with the active site present at the interface of two monomers (Kimber and Pai, 2000). The crystal structure of a CA from *E. coli* has also been resolved and has differences in the C-terminal structure to that of the pea enzyme, which have been proposed to influence the oligomeric state of the CA dimers (Cronk *et al.,* 2001). Like maize CA, the CA from the red alga *Porphyridium purpureum* has an internally repeating sequence, and the protein has two zinc molecules per monomer (Mitsuhashi *et al.,* 2000).

QuickTime™ and a
decompressor
are needed to see this picture.

Fig. 6.1. *Ribbon diagram of the crystal structure of the pea CA monomer, from N-terminus (blue) to C-terminus (red), with the zinc atom in the active site indicated by a blue sphere (Kimber and Pai, 2000).*

QuickTime™ and a
decompressor
are needed to see this picture.

Fig. 6.2. *Ribbon diagram of the crystal structure of the* E. coli *CA monomer (Cronk* et al.*, 2001). Helices are shown in red, and beta-strands in yellow. The magenta sphere indicates the location of the zinc atom.*

QuickTime™ and a
decompressor
are needed to see this picture.

Fig. 6.3. *Ribbon diagram of the crystal structure of the* P. purpureum *CA (Mitsuhashi* et al.*, 2000). The N-terminal is marked and coloured blue, the C-terminal is also marked and coloured green. The zinc atoms are shown as red spheres.*

In maize, the deduced amino acid sequences of the three CA transcripts identified, CA1, CA2 and CA3, were predicted to form subunits of approximately 49 kDa, 59 kDa and 71 kDa respectively (Section 5.3.1, Chapter 5). However, when a leaf protein extract was analysed by Western blotting using antibodies raised against a leaf-purified CA, four proteins bands were visualised at 27 kDa, 28 kDa, 47 kDa and 52 kDa (Burnell and Ludwig, 1997). These did not correspond to the predicted sizes, and suggested that maize CA mRNA may only be partially translated, or be subject to post-translational modification. The quaternary structure of the 27 kDa and 47 kDa species were determined by chromatography (Burnell and Ludwig, 1997). Both had native molecular weights of approximately 180 kDa, implying that the 27 kDa CA isoform associated at least as a hexamer, while the 47 kDa isoform was most likely a tetramer. Neither of the purified proteins could be sequenced from the N-terminal end, which prevented identification of a possible cleavage site.

6.1.4 Rationale

As CA from C_4 monocot species remains relatively uncharacterized, the aim of this investigation was to determine the structure of the maize CA isozymes and predict the sub-cellular location. The deduced amino acid sequence of CA from maize was analyzed for the presence of a putative transit peptide, and for post-translational modification sites. The secondary structure was also predicted using secondary structure prediction servers and by CD spectroscopy, the results of which were analyzed in comparison with the resolved crystal structures of other plant CAs.

In addition, to further characterize the maize CA isozymes several strategies were employed to purify CA from leaf tissue in order to obtain protein sequence data. While the nucleotide sequence of maize CA is known, no protein sequence has been previously obtained due to blockage of the N-terminus preventing sequencing by Edman degradation (Burnell and Ludwig, 1997). This problem was to be overcome using mass spectrometry and the expected result was the elucidation of protein sequence, potential splice sites and possible post-translational modifications.

Materials and methods

6.2.1 Analysis of maize CA protein sequence

6.2.1.1 Identification of a putative chloroplast transit peptide in the CA2 gene

The amino acid sequence obtained from translation of Exon 1 of the CA2 gene (Appendix 6.1) was submitted to the ChloroP 1.1 Prediction Server (http://www.cbs.dtu.dk/services/ChloroP/), a neural network based method used to identify chloroplast transit peptides and their cleavage sites based on derivations of the probability values.

6.2.1.2 Analysis of CA hydrophobicity

The hydrophobicity of the CA amino acid sequence, specifically the Repeat B amino acid sequence (Appendix 6.2) was analyzed using protein analysis tools available on the ExPASy server (http://www.expasy.ch/cgi-bin/protscale.pl). Whether hydrophobic regions identified formed part of a trans-membrane domain was analysed using the DAS trans-membrane prediction server (http://www.sbc.su.se/~miklos/DAS/tmdas.cgi).

6.2.2 Analysis of CA (Repeat B) secondary structure

6.2.2.1 Secondary structure prediction based on the amino acid sequence

The secondary structure of Repeat B was predicted by the use of two different computational algorithms, PHD (Rost, 1996; www.predictprotein.org), and the protein secondary structure finder at Softberry (www.softberry.com). The consensus for assigning a secondary structure element to a particular residue was if both algorithms predicted that element. The amino acid sequence is shown in Appendix 6.2.

6.2.2.2 Analysis of secondary structure by circular dichroism spectroscopy

Purification of bacterially-expressed Repeat B was performed as described (Section 5.2.3, Chapter 5). The expressed recombinant protein was used to generate CD spectra as it demonstrated the highest activity when assayed (Section 5.3.2.1, Chapter 5). Before analysis, the purity of the protein was assessed by SDS-PAGE and Coomassie blue staining, as well as immuno-blotting (Section 5.2.4.1, Chapter 5). The sample was dialyzed against buffer (50 mM Tris-HCl pH 7.5, 200 mM NaCl, 5 mM $MgCl_2$) for 16 h at 4°C. The buffer components were selected to ensure protein stability. The CD spectra were recorded in buffer at concentrations of 0.5 $mg.ml^{-1}$ and 0.25 $mg.ml^{-1}$ in a 1.0 cm pathlength cell, using a JASCO J-715 spectropolarimeter at room temperature. The molar ellipticity was calculated using the formula:

$$[\theta]_{\mathrm{ME}} = \frac{\theta}{(10CL)}$$

where θ is the measured ellipticity in degrees, C is the molar concentration and L is the pathlength (1.0 cm) (Aravind and Prasad, 2005). The molar concentration was calculated using the protein concentration ($g.ml^{-1}$), and the molecular weight (M_r): $C = (1000 \times P_{conc})/M_r$. The M_r of Repeat B was predicted to be 30 kDa. Data generated were analyzed using online CD secondary structure prediction servers (K2D2; http://www.ogic.ca/projects/k2d2/).

6.2.2.3 Analysis of Repeat B by native PAGE

Non-continuous native PAGE was performed using a 15% (w/v) polyacrylamide gel in 0.1 M Tris-borate buffer, pH 8.3 with a 5% (w/v) stacking gel. Electrophoresis was performed at 100 V for 4 h. The gel was subsequently stained with Coomassie blue (Section 2.2.1, Chapter 2).

6.2.3 Analysis of the CA2 unique 276 bp insert sequence

6.2.3.1 Purification of maize leaf mRNA

Maize leaf mRNA was prepared and quantified as previously described (Section 2.1.4, Chapter 2).

6.2.3.2 Generation of cDNA

One µg of RNA was used to generate cDNA using the QuantiTect Reverse Transcription kit and protocol (QIAGEN).

6.2.3.3 PCR

CA2 was analysed by PCR using cDNA generated from maize leaf mRNA. Several PCR primer pairs were used to amplify over the 276 bp insert of the CA2 sequence (Fig. 6.4, Table 6.1). Reactions were performed in 20 µl reaction volumes using GoTaq polymerase and buffer (Promega), and the reactions cycled as described (Section 2.1.14, Chapter 2).

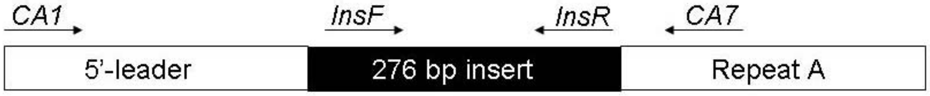


Fig. 6.4. *Schematic representation of primer annealing sites relative to the 276 bp insert of the CA2 cDNA sequence (gi:606810). The 5'-leader corresponds to the first exon, while the third exon encodes the start of Repeat A. Figure not to scale.*

Table 6.1. Sequence of the gene-specific primer pairs used for CA2 sequence analysis

Reaction	Primer 1	Sequence	Primer 2	Sequence
1	CA1 Fw	5′-atgtacacattgcccgtccgtg-3′	CA7 Rv	5′-ttcaagcgctcgacggt-3′
2	CA1 Fw	5′-atgtacacattgcccgtccgtg-3′	InsR	5′-gcccttggaggaagccttggaggg-3′
3	InsF	5′-cgccatggtctgtaaacgggacgg-3′	InsR	5′-gcccttggaggaagccttggaggg-3′

6.2.3.4 Sequence analysis

The PCR products were subcloned (Section 2.1.11, Chapter 2) for sequence analysis using the primer binding sites of the pGEM®-T vector system (Promega; Appendix 5.1), T7 and SP6. Sequence analysis was performed as described (Section 2.1.13, Chapter 2).

6.2.4 Analysis of CA from leaf tissue

Finely sliced leaf tissue (up to 32 g) was ground in liquid nitrogen and 1 ml.g^{-1} extraction buffer (50 mM Tris-HCl pH 8, 10 mM $MgSO_4$, 1 mM EDTA and 5 mM DTT). The homogenate was filtered through two layers of pre-wet Miracloth and centrifuged at 45,000 × *g* for 15 min at 4ºC. The concentration of protein in the supernatant (crude extract) was determined by Bradford assay using a standard curve generated with BSA (Section 2.2.5, Chapter 2). The crude extract was also assayed for CA activity (Section 5.2.4.4, Chapter 5).

6.2.5 Analysis of leaf proteins by Western blot

SDS-PAGE and Western blotting were performed as described in Section 2.2 (Chapter 2). The α-CA antibodies used for immuno-blotting were prepared and used as described in Section 5.2.4 (Chapter 5).

6.2.6 Gel filtration

A column (58 cm x 1.6 cm) of Sephacryl S-300 (Pharmacia) equilibrated with column buffer (50 mM Tris-HCl pH 8, 5 mM $MgCl_2$, 150 mM NaCl, and including 1% (v/v) Triton X-100 if proteins were extracted in the presence of detergent) was used to fractionate the proteins in a crude extract of maize leaves according to their native molecular weight. A sample that had been filtered through a 0.22 µm filter unit (Millipore), typically equivalent to 1.5% of the total column volume, was applied to the equilibrated column and fractions collected using a Pharmacia Biotech Frac-100 fraction collector. The flow rate was controlled with a Pharmacia Biotech peristaltic pump.

6.2.6.1 Column calibration

The Sephacryl S-300 column was calibrated using proteins of known molecular weights: ferritin (450 kDa, 100 $mg.ml^{-1}$, Sigma), lactate dehydrogenase (LDH, 140 kDa, 10 $mg.ml^{-1}$ Roche), malate dehydrogenase (MDH, 67 kDa, 10 $mg.ml^{-1}$, Roche) and cytochrome c (12.5 kDa, 5 $mg.ml^{-1}$, Sigma). A sample (equivalent to 1.5% total column volume) was prepared in column buffer and contained 10 µl ferritin, 15 µl LDH, 25 µl MDH and 60 µl of cytochrome c.

Protein peaks in the column eluate were determined by several methods. The relative A_{280} was used to determine in which fraction ferritin and cytochrome c had eluted, except in the presence of 1% (v/v) Triton X-100, where the relative A_{300} was used. The fractions containing LDH and MDH were determined by measuring enzyme activity spectrophotometrically following the oxidation of NADH. Reactions were performed in 125 mM Tris-HCl pH 7.5 with 0.2 mM NADH and 2 mM pyruvate (LDH) or oxaloacetate (MDH) and the rate of NADH oxidation was measured at 340 nm after addition of 50 µl of the eluted fraction.

The elution volume of each protein was calculated and used to determine the distribution coefficient (K_{dist}) according to the equation:

$$K_{dist} = (v_e - v_o)/(v_t - v_o)$$

Where v_e is the volume of the peak height, v_o is the void volume and v_t is the volume of the total packed bed. The K_{dist} was then used to generate a standard curve.

6.2.6.2 Identification of fractions containing CA

Several methods were used to determine in which fraction CA had eluted. These included measuring the CA activity in each eluted fraction, which was performed as described in Section 5.2.4.4 (Chapter 5), or by Western blotting of eluted fractions (Section 2.2, Chapter 2). Alternatively, the fraction in which CA had eluted was determined using a self-generated indirect enzyme linked immuno-sorbent assay (ELISA).

6.2.6.3 ELISA

The 96-well plates (Greiner Bio-One, flat bottom, medium binding) were first coated with 50 µl of each eluted fraction, in duplicate, for 12 – 16 h at 4ºC. The plates were washed in TBST (25 mM Tris-HCl pH 8, 150 mM NaCl, 0.1% v/v Tween-20) three times, and then blocked with 5% (w/v) skim milk powder in TBST using 200 µl per well for 2 h at 25ºC. The plates were again washed three times before addition of 50 µl of primary antibody (blot-affinity purified at 1:500 dilution, or polyclonal at 1:2,000 dilution; Section 5.2.4, Chapter 5) in TBS, and incubated for 4 h at 25ºC. The plates were washed six times and 50 µl of the secondary antibody (goat anti-rabbit IgG(H+L)-HRP conjugate, BioRad) was added at 1:1,000 dilution in TBS for 2 h. The plates were washed six times and 50 µl of 3,3',5,5'-Tetramethylbenzidine Liquid Substrate System (TMB, Sigma) added. When the reaction had proceeded sufficiently (as observed by colour change), 50 µl 0.5 M sulphuric acid was added to stop reactions and the A_{450} determined.

To ensure that the signal generated was not due to non-specific primary antibody interactions a competition assay was performed. A positive sample that had been diluted 1:10 with CA buffer (150 mM NaCl, 50 mM Tris-HCl pH 8, 5 mM $MgCl_2$) was used to coat an ELISA plate for 12 – 16 h at 4ºC. The same sample was diluted (1:10, 1:20, 1:50, 1:100, 1:500) and incubated with α-CA (Section 5.2.4, Chapter 5) for 12 – 16 h at 4ºC. The plate was washed three times with TBST and blocked with 5% (w/v) skim milk powder for 2 h. The plate was again washed three times with TBST and incubated with 50 µl of the primary antibody for 4 h. The diluted protein/antibody mixtures were used as the source of primary antibody. The plate was washed six times with TBST and incubated with 50 µl of the secondary antibody (goat anti-rabbit IgG(H+L)-HRP conjugate) added at 1:1,000 dilution in TBS for 2 h. The plate was washed six times and 50 µl TMB added. When the reaction had proceeded sufficiently (as observed by a colour change), 50 µl 0.5 M sulphuric acid was added to stop reactions and the A_{450} determined.

6.2.6.4 Trichloroacetic acid precipitation

In some cases, protein in eluted fractions was undetectable. Attempts were made to concentrate the protein by trichloroacetic acid (TCA) precipitation. TCA was added to each protein sample (30% w/v in a 1:1 ratio), and incubated on ice for 10 min. After centrifugation at 16,100 × *g* for 15 min, the supernatant was discarded and 1 ml of cold ethanol with 40 mM acetic acid was added to the pellet. This was again centrifuged and 1 ml of cold acetone was added to the pellet. After centrifugation at 16,100 × *g* for 15 min, the pellet was dried then resuspended in cracking buffer (0.01% w/v bromophenol blue, 125 mM Tris-HCl pH 6.7, 2% w/v SDS, 10% v/v glycerol, 10% v/v 2-mercaptoethanol) and incubated at 95ºC for 10 min. The samples were then analyzed by SDS-PAGE and Western blotting (Section 5.2.4, Chapter 5).

6.2.7 Ammonium sulphate precipitation

After preparation of a crude extract of leaf proteins, the sample was precipitated with ammonium sulphate. The volume of the sample was measured, and finely ground ammonium sulphate was added slowly with mixing at 4ºC. Precipitation was performed initially to 30%, then subsequently to 45%, 55% and 65% saturation. Once the ammonium sulphate was dissolved, the sample was allowed to mix for a further 30 min at 4ºC before centrifugation at 20,000 rpm for 20 min at 4ºC using a Beckman JA25.5. After centrifugation, the pellet was resuspended in 0.1 M Tris-HCl pH 7.5 and dialyzed for 16 h at 4ºC in preparation for ion exchange chromatography.

6.2.8 Ion exchange chromatography

A column (10 cm x 2.5 cm) of diethylaminoethyl (DEAE) Sepharose (GE Life Sciences) was prepared and equilibrated with 0.1 M Tris-HCl pH 7.4 (and including 1% v/v Triton X-100 if proteins were extracted in the presence of detergent). The resuspended and dialyzed fractions obtained from ammonium sulphate precipitation that displayed measurable CA activity (Section 5.2.4.4, Chapter 5) were applied to the column, and 100 ml fractions were collected. The column was washed with 50 mM Tris-HCl pH 7.4 containing 0.1 M NaCl, and proteins eluted sequentially using 0.2 M, 0.5 M and 1.0 M NaCl.

Proteins eluted from the column were analyzed by SDS-PAGE and Western blotting (Section 2.2, Chapter 2). The α-CA antibodies used for immuno-blotting were prepared and used as described in Section 5.2.4 (Chapter 5). The activity of CA in eluted fractions was measured as described in Section 5.2.4.4 (Chapter 5).

6.2.9 Immuno-precipitation

A crude extract of leaf proteins was prepared by finely grinding 5 g of maize leaf tissue in 10 ml buffer (50 mM HEPES-KOH pH 7.5, 10 mM $MgSO_4$, 10 mM DTT). The preparation was then filtered through two layers of pre-wet Miracloth and centrifuged at 20,000 rpm for 25 min at 4ºC using a Beckman JA25.5 rotor. The supernatant was filtered through a 0.45 µm filter and used for immuno-precipitation. Reactions included 0.1 ml Protein A-Agarose (Pierce), 1 ml of the crude protein extract and 10 µl α-CA or 50 µl blot-affinity purified antibodies (Section 5.2.4.2, Chapter 5). The reactions were incubated for 1 h at 4ºC, and then centrifuged at 300 × *g* for 1 min. The pellets were washed three times with TBST, and the final supernatant was prepared for analysis by SDS-PAGE and Western blotting by the addition of an equal volume of cracking buffer and incubation at 95ºC for 5-10 min. The pellets were resuspended in 100 µl cracking buffer and incubated at 95ºC for 5-10 min. SDS-PAGE and Western blotting were performed as described in Section 2.2 (Chapter 2).

Results

6.3.1 Analysis of the CA2 gene sequence

6.3.1.1 Identification of a putative chloroplast transit peptide

The first exon of the CA2 gene may translate and function as a chloroplast transit peptide as predicted by the ChloroP 1.1 Prediction Server. The probability that each amino acid forms part of a chloroplast transit peptide in relation to the other amino acids in the sequence was determined (Fig. 6.5).

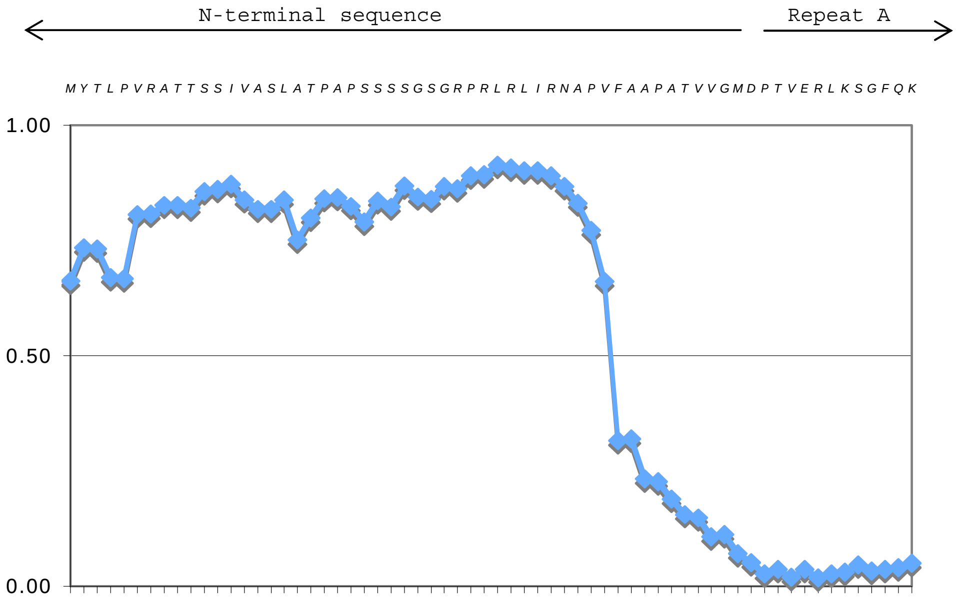


Fig. 6.5. *Probability of the translated sequence of Exon 1 (N-terminal sequence) acting as a chloroplast transit peptide.*

In rice, CA was found to have a chloroplast transit peptide of 63 amino acids determined by N-terminal sequencing of the purified mature protein (Suzuki and Burnell, 1995). When aligned with the translated sequence of Exon 1, despite being 21 amino acids shorter, the cleavage site as well as much of the transit peptide sequence is conserved (ChloroP 1.1 Prediction Server; Fig. 6.6).

```
Rice                MSTAAAAAAAQSWCFATVTPRSRAT---
VVASLASPSPSSSSSSSSNSSNLPAPFRPRLIRNT
Maize        MYT---------------LPV-RATTSSIVASLATPAPSSSSGSGRP--------RLRLIRNA


                ↓
Rice         PVF AAPVAPAAMDAAVDRLKDGFAKFKTEFYDKKPELFEPLKAGQAPKYMVFSCADSRVC
Maize        PVF AAPATVVGMDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVC
```

Fig. 6.6. *Alignment of the rice and maize CA chloroplast transit peptide amino acid sequences. The putative cleavage site of the chloroplastic transit peptide is indicated with an arrow.*

The rice sequence encodes a chloroplastic CA, with significant homology to the maize CA amino acid sequence (Fig. 6.7).

```
Maize        MDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVCPSVTLGLQPGE
Rice         MDAAVDRLKDGFAKFKTEFYDKKPELFEPLKAGQAPKYMVFSCADSRVCPSVTMGLEPGE
             **.:*:***.** *****.************:**:*:****:*:*********:**:***

Maize        AFTVRNIASMVPPYDKIKYAGTGSAIEYAVCALKVQVIVVIGHSCCGGIRALLSLKDGAP
Rice         AFTVRNIANMVPAYCKIKHAGVGSAIEYAVCALKVELIVVIGHSRCGGIKALLSLKDGAP
             ********.***.* ***:**.*************::******* ****:**********

Maize        DNFTFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEGLAG
Rice         DSFHFVEDWVRTGFPAKKKVQTEHASLPFDDQCAILEKEAVNQSLENLKTYPFVKEGIAN
             *.* ******* * ***:**:.****:******:******** **:***:*******:*.

Maize        GTLKLVGAHSHFVKGQFVTWEP
Rice         GTLKLVGGHYDFVSGNLDLWEP
             *******.* .**.*::  ***
```

Fig. 6.7. *ClustalW alignment of Repeat A of the maize CA sequence and the rice chloroplastic CA (not including the chloroplast transit peptide).*

6.3.1.2 Analysis of Hydrophobicity

A hydrophobicity plot for the amino acid sequence of Repeat B was generated (Kyte and Doolittle, 1982; Fig. 6.8).

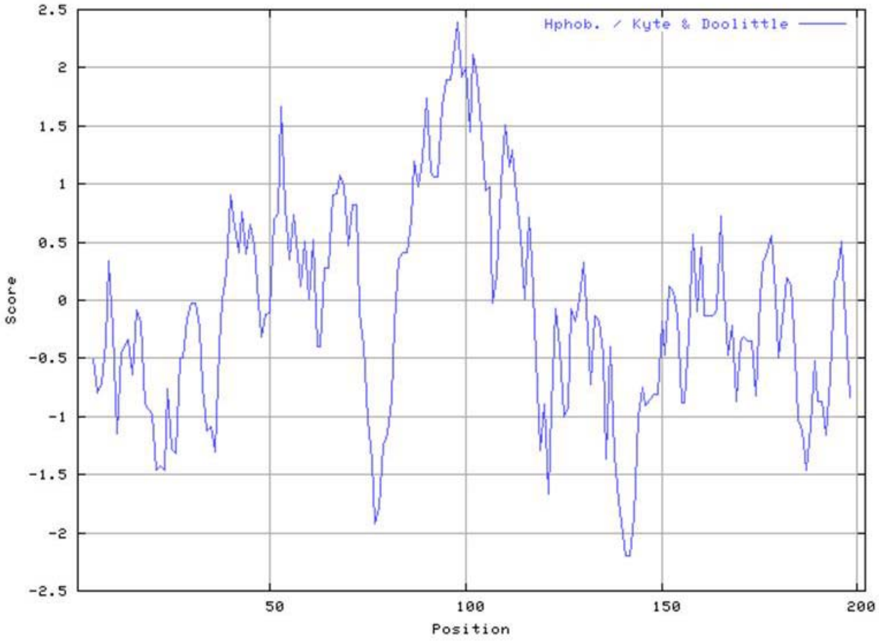


Fig. 6.8. *Hydrophobicity plot based on the amino acid sequence of Repeat B (Appendix 6.2).*

The sequence around position 100, which is halfway through the Repeat, is hydrophobic and may represent a membrane domain (Fig. 6.8). The sequence, *AIEYAVCALKEVLVVIGH*, may form part of a trans-membrane domain as predicted by the PHD prediction server and by analysis of potential trans-membrane domains by the DAS prediction server (Fig. 6.9).

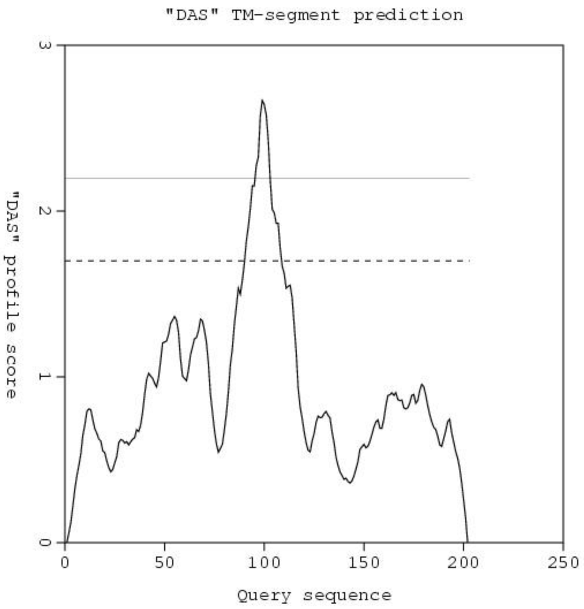


Fig. 6.9. *Graphical output from the DAS trans-membrane prediction server. The full line represents a strict cutoff, while the dotted line represents a loose cutoff.*

6.3.2 Analysis of CA (Repeat B) secondary structure

6.3.2.1 Secondary structure prediction based on the amino acid sequence

The secondary structure of Repeat B was predicted using two different computational algorithms. The secondary structure was found to be predominantly alpha-helical (H), with some beta-sheet (b) characteristics (Fig. 6.10). The secondary structure was assigned only if both algorithms predicted that element.

```
                   10        20        30        40        50
           MPQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMVFACSDSRVC
PHD          HHHHHHHHHH HHHHHH     HHHHHHHH     bbbbbbb
SBer         HHHHHHHHH   bbbbbbbb               bbbbbbbb


                   60        70        80        90       100
           PSVTLGLQPGEAFTVRNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLV
PHD          bbbb     bbbbbbb              HHHHHHHHHHH    bbb
SBer        bbbbbb       HHHHHHH              HHHHHHHHH   bbb


                  110       120       130       140       150
           VIGHSCCGGIRALLSLKDGAPDNFHFVEDWVRIGSPAKNKVKKEHASVPF
PHD        bbb    HHHHHH            HHHHHHHHHHHHHHHHHHH     H
SBer       bb       HHHHHHHH   HHHbbbbbbbbbb                H


                  160       170       180       190       200
           DDQCSILEKEAVNVSLQNLKSYPFVKEGLAGGTLKLVGAHYDFVKGQFVT
PHD        HHHHHHHHHHHHHHHHHH     HHHHHHH   bbbbbbbbbbb  bbbb
SBer       HHHHHHHHHHHHHHHHHHHHH           bbbbbbbbbb


           WEP
PHD        b
SBer
```

Fig. 6.10. *Secondary structure predictions based on the amino acid sequence of Repeat B. Alpha-helices are indicated (H), and beta-sheets (b). The prediction based on use of the PHD server is indicated by 'PHD' and the Softberry secondary structure prediction by 'Sber'. These are shown in red where both algorithms have predicted the same secondary structure.*

6.3.2.2 Analysis of secondary structure by CD spectroscopy

Purified and bacterially expressed Repeat B was analysed using CD spectroscopy. The purity of Repeat B was assessed by SDS-PAGE and Western blotting with antibodies against the N-terminal histidine tag generated when using the pROEx^{TM}-HT vector (Fig. 6.11; Appendix 5.1 for vector diagram).

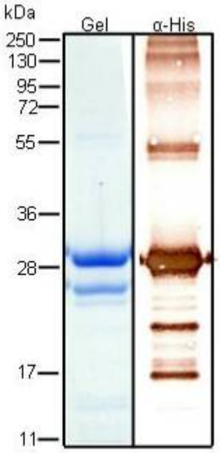


Fig. 6.11. *Coomassie blue-stained gel and corresponding Western blot of the Repeat B sample used for circular dichroism analysis. 5 µl of protein was loaded per lane, equivalent to 7.5 µg of total protein (1.5 mg.ml^{-1} protein concentration).*

The data generated when 0.25 mg.ml^{-1} protein was analysed were subjected to the online CD secondary structure prediction server (K2D2; http://www.ogic.ca/projects/k2d2/). The error was significant, which may be due to both the buffer used and the presence of contaminating proteins that were not removed during purification (Fig. 6.11). Despite this, the resulting prediction is for less than 2% beta sheet and at least 82% alpha-helical characteristics (Fig 6.12).

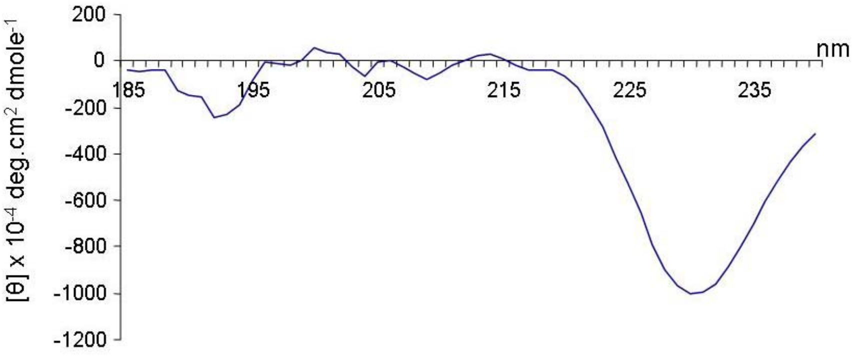


Fig. 6.12. *CD spectra for Repeat B*

6.3.2.3 Analysis of Repeat B by native PAGE

The quaternary structure of Repeat B was also investigated by gel filtration (results not shown) and native-PAGE. Under native conditions Repeat B migrated as a dimer at approximately 55 kDa (Fig. 6.13).

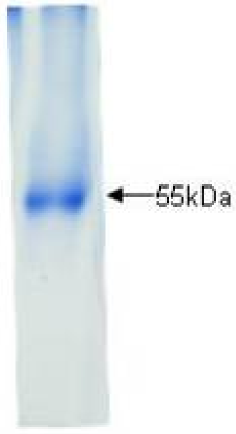


Fig. 6.13. *Coomassie blue-stained discontinuous native PAGE of Repeat B (7.5 µg total protein).*

6.3.3 Analysis of the CA2 unique 276 bp insert sequence

Sequencing data around the 276 bp insert of CA2 was obtained using cDNA generated from maize leaf mRNA as template (Section 5.2.1, Chapter 5). The primers used in these reactions were designed to amplify over the 276 bp insert (Table 6.1). The amplification products were subcloned and the sequence analyzed (Fig. 6.14). Analysis of the sequence data generated indicated that there were two changes in the reading frame of this component of the CA2 transcript.

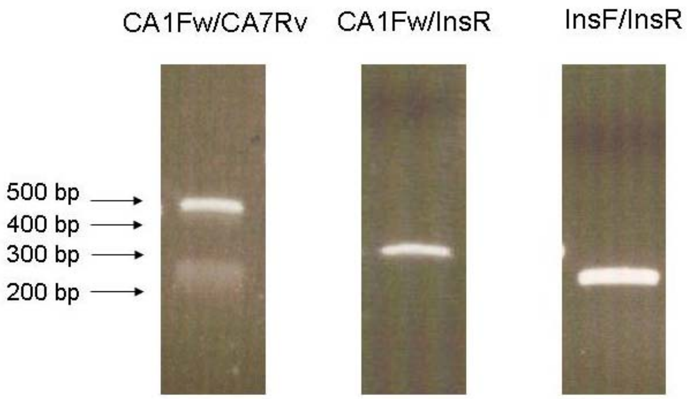


Fig. 6.14. *PCR products generated from the primer combinations indicated in Table 6.1. The DNA fragments shown are representative of a number of repetitions of these reactions.*

The first change in the reading frame observed generated a stop codon at the start of the second exon (Fig. 6.15). This was confirmed by alignment with genomic sequence at this position. The actual sequence in the same reading frame as the initiating codon of the cDNA sequence (ATG) places a stop codon (TAA) at the start of the second exon, with the first nucleotide of the first codon (C) coming from the 3′-end of the first exon.

```
AZM4_68974  CGTCATCATGTTGCT AGT AAA CGG GAC GGC GGG CAG CTG AGG AGT CAA ACG AGA GAG
                             || ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
CA2 cDNA                    TGT AAA CGG GAC GGC GGG CAG CTG AGG AGT CAA ACG AGA GAG
                             C   K   R   D   G   G   Q   L   R   S   Q   T   R   E

Observed reading frame     CTG TAA ACG
                            L   *
```

Fig. 6.15. *Alignment of the genomic sequence (AZM4_68974) with the CA2 sequence (as reported, accession gi:606810) at the start of the second exon. The reported (CA2 cDNA) and the observed reading frames are presented, with the stop codon indicated with an asterisk.*

The second change was an additional C nucleotide at position 241 in the alignment shown in Fig. 6.16. This nucleotide is not present in the reported sequence (NCBI database, gi:606810; Burnell and Ludwig, 1997), but was identified in the cDNA sequences generated as well as genomic DNA sequence from the library screen and in the genomic DNA assembly AZM4_68974 (Section 3.3.2.3, Chapter 3; Appendix 3.3B and 3.4D).

```
                                10           20           30
gi606810         TGT AAA CGG GAC GGC GGG CAG CTG AGG AGT CAA ACG AGA
cDNA            CTGT AAA CGG GAC GGC GGG CAG CTG AGG AGT CAA ACG AGA
                  C   K   R   D   G   G   Q   L   R   S   Q   T   R

                 40           50           60            70
gi606810         GAG ATC GAG AGA GAA AGA AAG GGA GGG CAT CCA CCA GCC
cDNA             GAG ATC GAG AGA GAA AGA AAG GGA GGG CAT CCA CCA GCC
                  E   I   E   R   E   R   K   G   G   H   P   P   A

                  80           90            100          110
gi606810         GGC GGG CAT AAG AGG GGA GGA GAG AGA GGC CAG AGA AGA
cDNA             GGC GGC GAT AAG AGG GGA GGA GAG AGA GGC CAG AGA AGA
                  G   G   H   K   R   G   G   E   R   G   G   R   R

                   120           130          140          150
gi606810         GGA GGA GAA GAA GAA GAA GAT GAG CAG CTG CCT CTG CCT
cDNA             GGA GGA GAA GAA GAA GAA GAT GAG CAG CTG CCT CTG CCT
                  G   G   E   E   E   E   D   E   Q   L   P   L   P

                     160          170          180           190
gi606810         TCC GAA AAA AAA GGA GGG GCC AGC GAA GGA GAA GCC GTC
cDNA             TCC GAA AAA AAA GGA GGG GCC AGC GAA GGA GAA GCC GTC
                  S   E   K   K   G   G   A   S   E   G   E   A   V

                      200          210           220          230
gi606810         CAC AGA TAC CCC CAC CTC GTC ACT CCT TCA GAA CCA GAA
cDNA             CAC AGA TAC CCC CAC CTC GTC ACT CCT TCA GAA CCA GAA
                  H   R   Y   P   H   L   V   T   P   S   E   P   E

                       240            250          260          270
gi606810         GCC CTC C-A A CCT CCA CCT CCT CCC TCC AAG GCT TCC TCC
cDNA             GCC CTC CCA A CCT CCA CCT CCT CCC TCC AAG GCT TCC TCC
                  A   L   Q     P   P   P   P   P   S   K   A   S   S

                       279
gi606810         AAG GGC
cDNA             AAG GGC
                  K   G
```

Fig. 6.16. *Alignment of the reported CA2 sequence (gi:606810) with the sequence obtained from analysis of maize leaf mRNA. Numbering is from the start of the second exon representing the 276 bp insert unique to CA2.*

6.3.4 Analysis of CA from leaf tissue

While the aim was to purify CA, no method used enabled purification of maize leaf CA to homogeneity.

6.3.4.1 Analysis of maize leaf proteins by Western blot

The preparation of proteins extracted from maize leaf (crude extract) was analyzed by Western blotting using α-CA (Fig. 6.17). Four immuno-reactive proteins were detected at 27 kDa, 28 kDa, 47 kDa and 52 kDa, in agreement with that previously reported (Burnell and Ludwig, 1997).

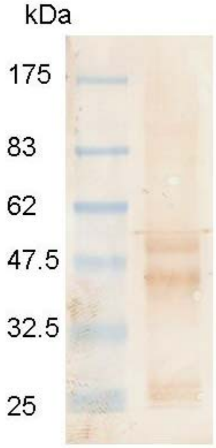


Fig. 6.17. *Western blot of 10 μl of the preparation of maize leaf proteins (1.3 mg.ml^{-1}) using α-CA antibodies.*

6.3.4.2 Gel filtration

To determine the native molecular weight of CA present in maize leaves, proteins extracted from leaf tissue were analyzed by gel filtration. This included analysis in the presence of the detergent Triton X-100 to ensure any membrane bound proteins were also isolated, however these analyses were unsuccessful (results not shown).

A Sephacryl S-300 column was used to separate proteins extracted from maize leaf tissue. To determine the native molecular weight of the eluted proteins, the column was calibrated using proteins of known molecular weight; ferritin (440 kDa), LDH (140 kDa), MDH (67 kDa) and cytochrome c (12.5 kDa; Fig. 6.18). The fractions in which ferritin and cytochrome c eluted was determined by measuring the relative absorbance at 280 nm, whereas the fractions in which LDH and MDH had eluted was determined by measuring dehydrogenase activity by the oxidation of NADH at 340 nm (Section 6.2.6.1).

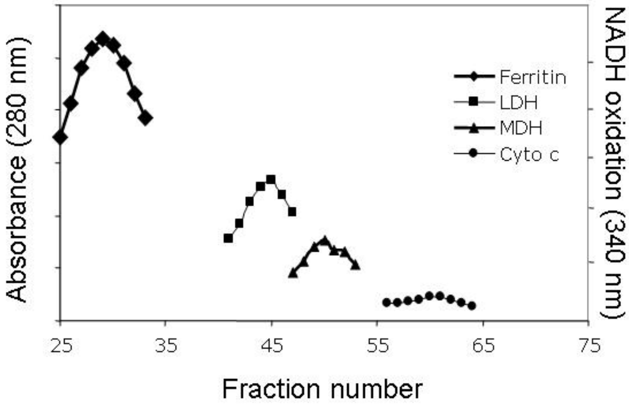


Fig. 6.18. *Typical elution profile of the proteins used for column calibration. The absorbance at 280 nm was used to determine in which fractions ferritin and cytochrome c (Cyto c) had eluted, while the rate of NADH oxidation measured at 340 nm was used to determine in which fractions LDH and MDH had eluted.*

From these results, the eluted volume was determined (v_e), which enabled the distribution coefficient (K_{dist}) to be calculated and a standard curve generated (Fig. 6.19). The equation to the standard curve could then be used to determine the native molecular weight of unknown proteins.

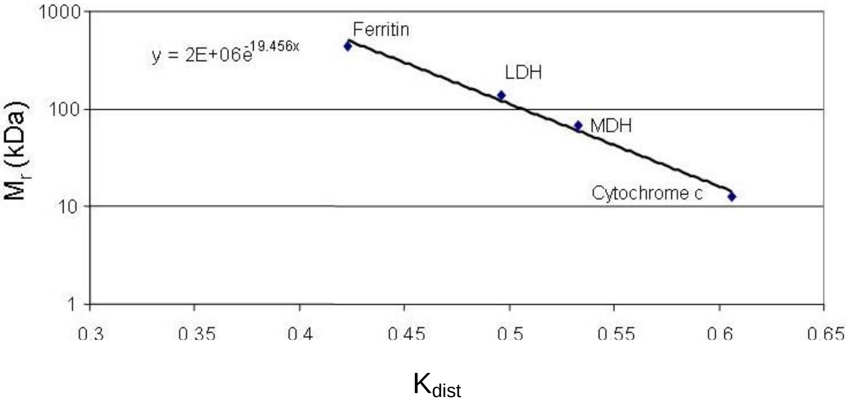


Fig. 6.19. *Typical standard curve relating native molecular weight (M_r) to the distribution coefficient (K_{dist}) for four known protein samples, with the equation to the line indicated.*

After gel filtration, CA was not detected in the eluted fractions, whether by activity assay (Section 5.2.4.5, Chapter 5) or by Western blotting with α-CA antibodies. Several techniques were employed in order to increase sensitivity. The first of these was the creation of an ELISA (Section 6.2.6.3). A sample of each fraction that had eluted from the gel filtration column was analyzed, and it was possible to detect peaks that corresponded to CA-containing fractions (Fig. 6.20).

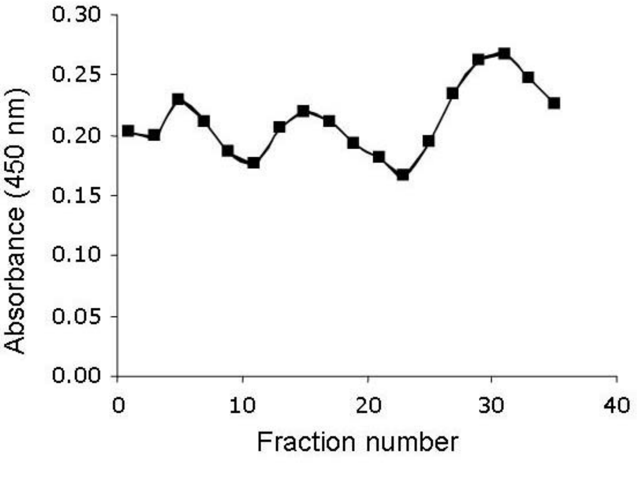


Fig. 6.20. *Detection of CA in fractions eluted from the gel filtration column by ELISA.*

Peaks of antigenic proteins (CA) were detected in fractions 7, 15 to 17, and 29 to 31 (Fig. 6.20). By determining the elution volume (v_e) of these peaks, and using the standard curve generated, the size ranges of these peaks were calculated (Table 6.2). The three peaks correspond to native molecular weights of approximately 600 kDa, 150-200 kDa and 22-29 kDa, respectively. The 150-200 kDa peak corresponded to the 180 kDa CA isoforms previously identified in maize leaf (Burnell and Ludwig, 1997). The presence of a peak at approximately 22-29 kDa may correlate with a CA monomer (Fig. 6.17).

Table 6.2. The estimated molecule weight of proteins in the peaks observed (Fig. 6.20), determined using the standard curve generated (Fig. 6.19).

Fraction number	**v_e (ml)**	**K_{dist}**	**M_r (kDa)**
7	60.13	0.29	612.4
15	64.85	0.35	203.2
16	65.44	0.36	176.9
17	66.03	0.37	154.2
29	73.11	0.45	29.5
30	73.70	0.46	25.7
31	74.29	0.47	22.4

To confirm these results, these specific fractions were analyzed by Western blotting, however no protein could be detected (results not shown). The protein in each fraction was then concentrated by TCA precipitation (Section 6.2.6.4), and again analyzed by Western blotting. Only one band at approximately 52 kDa could be detected using α-CA antibodies in Fraction 7 (Fig. 6.21). There were no bands obvious in the remaining fractions (data not shown).

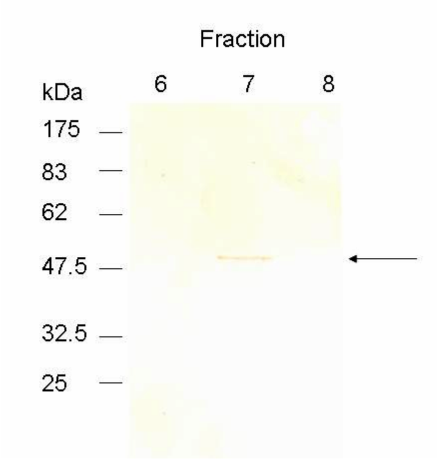


Fig. 6.21. *Western blot showing a 52 kDa immuno-reactive band in Fraction 7 after concentration by TCA precipitation.*

6.3.4.3 Ammonium sulphate precipitation and ion exchange chromatography

Up to 20 g of leaf tissue was used to extract maize leaf proteins. After precipitation with ammonium sulphate, CA-containing fractions were subjected to ion exchange chromatography. This was also performed in the presence of the detergent Triton X-100 to ensure isolation of any membrane-associated leaf proteins. While separation was achieved (Fig. 6.22), CA was not purified to homogeneity.

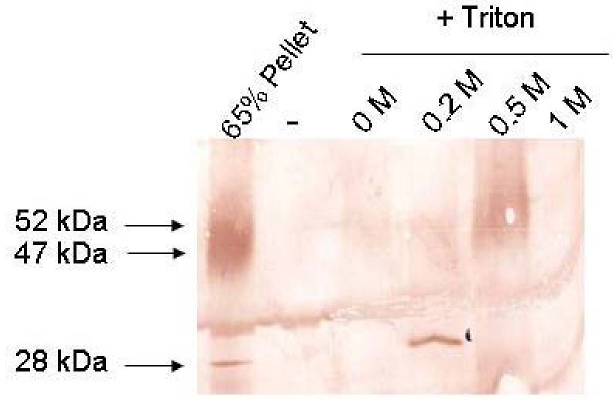


Fig. 6.22. *Analysis of CA by Western blotting after precipitation with ammonium sulphate (65% pellet), and analysis of eluted fractions (0 M, 0.2 M, 0.5 M and 1 M NaCl) obtained after subsequent separation by ion exchange chromatography in the presence of Triton X-100.*

6.3.4.4 Immuno-precipitation

The α-CA antibody used for this analysis was unsuitable for immuno-precipitation experiments, which were unsuccessful (results not shown).

Discussion

6.4.1 Analysis of amino acid sequence data

6.4.1.1 Prediction of sub-cellular location

It has been hypothesized that individual exons encode protein domains with discrete functions, as demonstrated by the primary structure of the genes encoding the small subunit of Rubisco and PPDK in maize. In these cases, the first exon encodes a transit peptide directing sub-cellular location of the expressed protein (Matsuoka, 1990). The translated sequence of the first exon of the CA2 gene (Appendix 6.1) was analyzed using the ChloroP 1.1 Prediction Server, which indicated that there was a high probability that the amino acids composing this sequence form part of a chloroplast transit peptide (Fig. 6.5). Additionally, this sequence had considerable similarity to the rice CA chloroplast transit peptide with conservation of the cleavage site (Fig. 6.6).

CA from *F. bidentis,* which is also a C_4 species, has a transit peptide (Cavallaro *et al.,* 1994). Western blotting of the crude leaf protein extract showed that the expressed protein had a subunit size similar to the predicted molecular weight based on the amino acid sequence including the transit peptide. Therefore, it was concluded that this CA was not processed and remained in the cytosol (Tetu *et al.,* 2007). CA from maize leaf has been purified to homogeneity, although blockage of the N-terminus prevented sequence data from being obtained and identification of a possible cleavage site (Burnell and Ludwig, 1997). The amino acid sequence of the *F. bidentis* transit peptide displays similarity to the rice and maize CA transit peptides, including conservation of the cleavage site (Fig. 6.23).

```
Rice          MSTAAAAAAAQSWCFATVTPRSRATVVASLASPSPSSSSSSSSNSSNLPAPFRPRLIRNT
Maize         MYTLPVRATTS-------------SIVASLATPAPSSSSGSGRP--------RLRLIRNA
F.bidentis    MSAASAFAMNAPSFNASSLKKASTSARSGLSARFTCNSSSSSSSSS----ATPPSLIRNE
              * : .. *                :  :.*::   ...**.*.             ****

                   ↓
Rice          PVFAA PVAPAAMDAAVDRLKDGFAKFKTEFYDKKPELFEPLKAGQAPKYMVFSCADSRVC
Maize         PVFAA PATVVGMDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVC
F.bidentis    LVFAA PAPIITPN----WTEDGNESYEEAIDALKKTLIE---KGEL--------------
               *****.. :   .:* .:: .  * *:*  *:
```

Fig. 6.23. *Alignment of the rice, maize and F.bidentis CA putative transit peptide sequences. The cleavage site of the chloroplast transit peptide is indicated with an arrow.*

In vivo chloroplast uptake assays of the three CAs from *F. bidentis* indicated that only one of the isozymes was transported into the chloroplast, despite the presence of the transit peptide on all three (Tetu *et al.,* 2007). As both *F. bidentis* and maize are C_4 plants and have similar carbon requirements, it is likely that of the three CA isozymes in maize, one is also imported into chloroplasts. Amplification of the 5′-leader sequence in semi-quantitative reverse transcriptase-PCR experiments showed that transcription of this part of the maize CA transcript was occurring at relatively low levels along the full length of the maize leaf (Section 4.3.4, Chapter 4).

6.4.1.2 Hydrophobicity

The hydrophobicity plot of an amino acid sequence allows for identification of any hydrophobic regions as well as membrane spanning domains (Kyte and Doolittle, 1982). A hydrophobicity plot for the amino acid sequence of Repeat B was generated in order to further characterize the expressed protein based on the deduced amino acid sequence (Fig. 6.8; Appendix 6.2).

There was a region of hydrophobicity around position 100 in the amino acid sequence, which is midway through the repeating region, and could potentially correspond to a membrane spanning domain. This was confirmed by analyzing the sequence using the DAS prediction server with the amino acid sequence *AIEYAVCALKEVLVVIGH* likely to form part of a trans-membrane domain (Fig. 6.9). However, this region of the protein contains several of the residues that have been identified in beta-CAs from other plant species as being involved in catalysis, specifically those amino acids that are potential ligands of the catalytic zinc molecule. In the sequence shown above, these are the first glutamate and the last histidine residue. Glutamate was identified as a potential zinc ligand, while the histidine residue has been mutated in pea CA causing a decreased ability to bind zinc (Bracey *et al.,* 1994; Cronk *et al.,* 2001). The crystal structure of the pea CA has shown that the catalytic site is at the interface of two monomers creating a hydrophobic pocket in which the catalytic zinc ion residues (Kimber and Pai, 2000). It is most likely that this hydrophobic region of the protein actually represents the active site of the maize CA isozymes.

A CA isozyme has been associated with the plasma membrane of the mesophyll cells of C_4 plants, but is unlikely to be an integral membrane protein (Utsunomiya and Muto, 1993; Ivanov *et al.,* 2007). An increase in the amount of the 28 kDa CA isozyme recovered after treatment with a detergent was observed (Burnell and Ludwig, 1997). It was proposed that this CA facilitates transport of carbon dioxide across the membrane, and along with the cytoplasmic CA, supplies bicarbonate for the reaction catalyzed by PEP carboxylase. Furthermore, the plasma membrane associated CA may represent between 20 to 60% of total CA activity detectable (Ivanov *et al.,* 2007). Bacterially expressed Repeat B demonstrated the highest CA activity (Section 5.3.2.1, Chapter 5), and it may be that this isozyme represents the plasma membrane associated CA.

6.4.1.3 Secondary structure predictions

Predictions of secondary structure using computer algorithms as well as protein analysis using techniques such as CD spectroscopy have been shown to be reliable methods (Lees and Janes, 2008). The predicted secondary structure of CA based on the deduced amino acid sequence of Repeat B was predominantly alpha-helical (Fig. 6.10; Appendix 6.2). This is in agreement with original predictions made from CD spectroscopy of CAs from other plant species, and with the CD spectra obtained in this analysis (Fig. 6.12; Johansson and Forsman, 1993; Aravind and Prasad, 2005). It also correlates to the crystal structures that have been resolved for other beta-CAs, particularly the pea and *E. coli* enzymes (Fig. 6.1, Fig. 6.2).

When analysed by native PAGE, Repeat B associated as a dimer at approximately 55 kDa (Fig. 6.13). A dimer would be the minimum necessary oligomerization state required to form an active enzyme complex. As shown by the crystal structure of the pea CA, at least two monomers are required to form an active enzyme complex by enabling formation of the hydrophobic active site pocket (Kimber and Pai, 2000).

6.4.2 Analysis of the unique 276 bp insert of CA2

Analysis of the sequence generated by amplification over the 276 bp insert unique to CA2 consistently showed two changes in the open reading frame, when the sequence was translated in the same reading frame as the initiating methionine. The first of these was the presence of a stop codon (TAA) at the start of the 276 bp insert, which corresponded to the start of the second exon in the CA2 gene (Fig. 6.14). The presence of the stop codon was confirmed by alignment of the cDNA sequence with the genomic DNA sequence, and analysis of the intron/exon boundary sequences of the first and second exons where the consensus intron sequence (*ag*) was not present before the second exon (Chapter 4.3.1, Chapter 4). The second change in reading frame observed was at position 241, where an additional C nucleotide was found in the cDNA as well as the genomic DNA sequences (Fig. 6.15; Section 3.4.1.1, Chapter 3).

These changes would prevent accurate translation of the CA2 transcript, unless translation was initiated from Exon 3, which corresponds to the beginning of Repeat A. While the 5′-end of the mRNA transcript encoding CA2 may not be translated, its purpose may be to regulate mRNA stability or translation, a mechanism to allow the plant to respond to environmental factors such as light, temperature or stress. The 5′- and 3′-untranslated regions of several maize genes play a role in increasing mRNA stability by favouring association with polyribosomes, in response to oxygen levels or high light conditions (Bailey-Serres and Dawe, 1996; Hansen *et al.,* 2001). There may be a light-responsive element in the untranslated region of the CA2 transcript, and CA mRNA levels increase when the leaf is illuminated (Burnell and Ludwig, 1997).

Alternatively, the 5′-end of the mRNA transcript may direct cell-specific expression of the transcript. As maize uses the C_4 photosynthetic pathway, which involves both mesophyll and bundle sheath cells in the processes of carbon assimilation and fixation, the expression of many associated proteins is directed to a specific cellular location. For example, the 5′-untranslated region (UTR) of the nuclear-encoded small subunit of Rubisco directed the accumulation of a green fluorescent protein and the corresponding transcript specifically to the bundle sheath cells of a transformed *F. bidentis* leaf, leading to the conclusion that regulation was at the level of mRNA stability, and therefore accumulation occurred where the transcript was not degraded (Patel *et al.,* 2006).

CA expression occurs predominantly in the mesophyll cell as evidenced by localization of CA transcripts and enzyme activity (Poincelot, 1972; Burnell and Hatch, 1988; Wyrich *et al.,* 1998). When the relative abundance of the 276 bp insert region of CA2 was analyzed by semi-quantitative reverse transcriptase-PCR, the abundance of this part of the transcript increased towards the leaf tip in a similar way as that observed for PPDK (Section 4.3.4, Chapter 4). From this observation, it could be concluded that this part of the transcript is associated with the CA isoform involved in the operation of the C_4 pathway and therefore has a strict requirement to be located in the mesophyll cell cytoplasm. It has been previously hypothesized that the translated peptide may be associated with a membrane protein, as the deduced amino acid sequence of the 276 bp insert is particularly hydrophilic and unlikely to be associated directly with the membrane (Burnell and Ludwig, 1997; Utsunomiya and Muto, 1993). Whether this amino acid sequence forms part of the mature protein will only be determined when protein sequence data becomes available.

6.4.3 Analysis of CA from maize leaf

The aim of this analysis was to purify CA from maize leaf tissue in order to generate protein sequence data; however purification of CA to homogeneity was not achieved. Analysis of a crude extraction of maize leaf proteins, however, did confirm that four protein bands of approximately 27 kDa, 28 kDa, 47 kDa and 52 kDa were cross-reactive with CA antibody (Fig. 6.17), in agreement with that previously reported (Burnell and Ludwig, 1997).

The main difficulty experienced when trying to isolate CA was insufficient starting material (leaf), and therefore CA in extracted samples was difficult to detect. Several techniques were employed to aid detection including the creation of an ELISA and TCA precipitation of processed samples. After gel filtration, the use of an ELISA did enable three CA-containing fractions to be identified (Fig. 6.20), one of which was confirmed by Western blotting of TCA-precipitated samples (Fig. 6.21). These CA-containing fractions corresponded to CA with native molecular weights of approximately 600 kDa, 150-200 kDa and 22-29 kDa (Table 6.2). The identification of CA in fractions representing proteins ranging from 22-29 kDa most likely represented a CA monomer (27 kDa or 28 kDa), indicating that the gel filtration conditions were not suitable for maintaining the quaternary protein state. The immuno-reactive fractions corresponding to proteins with native molecular weights of 150-200 kDa may reflect separation of the 180 kDa CA isoforms previously identified by chromatography (Burnell and Ludwig, 1997). A 600 kDa CA isoform has not been previously identified in maize leaf tissue. This CA isoform appeared to represent an oligomer composed of the 52 kDa subunits identified by Western blotting of the crude maize leaf protein extraction (Fig. 6.21, Fig. 6.17; Burnell and Ludwig, 1997). Beta-CAs from dicot species have been reported as octamers, hexamers, tetramers and dimers (Hiltonen *et al.,*1998; Kimber and Pai, 2000).

Extraction of maize leaf CA was also performed in the presence of detergent (Triton X-100) to determine which CA isoform was associated with cellular membranes. The crude protein extraction was precipitated with ammonium sulphate, and CA-containing fractions were further separated by ion exchange chromatography. Separation of the different CA isoforms was achieved using this method, as could be detected by Western blotting (Fig. 6.22). However, whether these CA isoforms were retained in their native state was unlikely due to the high salt concentrations that were used to elute from the ion exchange column and evidenced by a lack of detectable CA activity in the eluted fractions.

Conclusion

The CA isozymes from many C_3 dicot plant species are well characterized, however the structure, sub-cellular location and function of CA from agronomically important C_4 monocot species remains largely unknown. These particular species are classified as NADP-malic enzyme (NADP-ME) type C_4 species, as they use NADP-ME to decarboxylate the four carbon acids that diffuse into the bundle sheath cells, releasing carbon dioxide in the vicinity of Rubisco and contributing to the efficiency of the C_4 pathway. Significantly, the CA transcripts from these species are distinct from other C_4 or monocot species, primarily due to transcript length (Burnell and Ludwig, 1997; Burnell, 2000).

The first exon of the CA2 gene was found to encode a chloroplast transit peptide, which had significant homology to the rice transit peptide. However, like the CA from the C_4 species *F. bidentis,* it is likely that only one of the three CA isozymes present in the leaf is transported into the chloroplast and that the transit peptide has simply been retained after evolution of the C_4 pathway, when CA location and function has changed. Future experiments to confirm this could be performed by using an *in vivo* chloroplast uptake system.

The secondary structure of Repeat B was confirmed by analysis of sequence data and CD spectra to be composed primarily of alpha-helices. This is in agreement with crystal structures obtained from other beta-CAs, such as pea, *E. coli* and *P. purpureum,* which also show alpha-helical characteristics (Fig. 6.1 - 6.3; Kimber and Pai, 2000; Cronk *et al.,* 2001; Mitsuhashi *et al.,* 2000). In addition, Repeat B was found to exist as a dimer by native PAGE, which would be the minimal oligomerization required for the enzyme to form a catalytic site. Ideally, the crystal structure of the maize CA isozymes could be resolved in order to confirm these results. However, the difficulty lies in obtaining leaf-purified enzyme. Several attempts were made at isolating the enzyme from maize leaf tissue, using techniques such as ammonium sulphate precipitation, ion exchange chromatography and gel filtration, but these were unsuccessful.

While the nucleotide sequences of the maize CA genes are publicly available, no protein sequence has been previously obtained due to the blockage of the N-terminus preventing sequencing by Edman degradation (Burnell and Ludwig, 1997). Several ambiguities remain including whether the 5′-leader sequence encoding the chloroplast transit peptide is retained on the mature enzyme and whether the 276 bp insert, which was found to contain several reading frame errors, is actually translated in the leaf. The amino acid sequence encoded by the 276 bp insert is particularly hydrophilic and may be suitable for the generation of antibodies that could be used for Western blotting or immuno-localization in the leaf.

CHAPTER 7

7. Conclusions

Maize CA is encoded by a multi-gene family resulting in the expression of several isozymes in the plant. The mRNA transcripts encoding these isozymes contain repeating sequences of approximately 600 bp that encode multiple protein domains (Repeat A, Repeat B and Repeat C). This characteristic is specific to NADP-malic enzyme type monocot C_4 species including agronomically important crops such as sorghum and sugarcane (Burnell and Ludwig, 1997; Wyrich *et al.,* 1998). In maize, three cDNA sequences have been determined and designated CA1, CA2 and CA3 (Burnell and Ludwig, 1997; Burnell, 2000).

There are at least three CA genes in the maize genome confirmed by Southern blotting and analysis of maize genomic sequence databases. The CA2 gene was isolated from a genomic DNA library and further characterization revealed that it encodes two identical protein domains, with two groups of six exons corresponding to the two repeating regions of the CA2 transcript. Apart from the CA2 gene, there are also at least two CA genes that only contain sequence encoding a single protein domain. Whether all of these genes are functional was not determined, but the complexity of the maize genome contributed to by the action of retro-transposons implies the likely presence of pseudo-genes as well as truncated and non-functional CA genes.

The second exon of the maize CA2 gene is transcribed but not translated and does not form part of the expressed protein. This sequence corresponds to a 276 bp insert that is unique to the CA2 transcript. This insert is not present in CA1, CA3, or the CA transcripts identified on the NCBI database, AY109272 and DQ246083. The nucleotide sequence of the 276 bp insert contained differences from the reported sequence (Genbank accession: U08401, version: gi:606810), which created errors in the reading frame that would result in attenuation of translation. However, the 276 bp insert does form part of the CA2 transcript, and encodes the CA isozyme that catalyzes the first reaction of the C_4 photosynthetic pathway, providing bicarbonate for PEP carboxylase in the cytoplasm of mesophyll cells. Semi-quantitative reverse transcriptase-PCR analysis showed that the relative abundance of this part of the transcript increased from the base of the leaf to the tip in a similar way as that observed for the transcript encoding the C_4 enzyme PPDK (Langdale *et al.,* 1988a; Langdale *et al.,* 1988b).

In maize CA is part of the C_4 photosynthetic pathway but is also involved in non-photosynthetic functions. Like CA from the C_4 dicot species, *Flaveria bidentis,* maize CA could be involved in lipid biosynthesis pathways and/or replenishing the Krebs cycle intermediates together with PEP carboxylase (Tetu *et al.,* 2007). In support of this, analysis of the CA2 gene sequence identified promoter elements that direct constitutive expression. In addition, CA transcripts were identified in root tissue by semi-quantitative reverse transcriptase-PCR.

When expressed as single domain proteins in a bacterial expression system, Repeat A and Repeat B were active, catalyzing the hydration of carbon dioxide and releasing hydrogen ions. The carbon dioxide hydration activity of Repeat B was relatively high compared to the activity of either Repeat A or C. Repeat B was also found to be a dimer when analyzed by native PAGE, and CD spectra indicated it was composed primarily of alpha-helices, in agreement with that observed for other plant CAs (Kimber and Pai, 2000).

The CA isozymes expressed in maize as single domain proteins would correlate with two protein bands of approximately 27 kDa and 28 kDa that were immuno-reactive with a plant monocot beta-CA probe (Burnell and Ludwig, 1997). Additionally, the 47 kDa protein band identified could correspond to CA expressed with two protein domains, for example CA1, but without translation of the 5′-leader sequence or the 276 bp insert of the CA2 transcript. When CA1 was bacterially expressed, it also exhibited CA activity. Ultimately protein sequence data, which was not obtained in this study, is required to correlate which part of the transcript is expressed and whether other post-transcriptional or post-translation modifications are occurring.

The active site of the individual protein domains, Repeat A, Repeat B and Repeat C was identified and found to contain the conserved amino acids proposed to coordinate the catalytic zinc ion and act as a proton acceptor during regeneration of the active enzyme complex (Kimber and Pai, 2000; Cronk *et al.,* 2001). Further experiments to change these amino acids, using site directed mutagenesis, could confirm the role these amino acids play in catalysis. The CA activity of Repeat C was relatively low, and several differences in the amino acid sequences of Repeat C compared to Repeat A or Repeat B were identified as contributing to the lack of activity.

This research has confirmed the earlier findings of the presence of repeating sequences in maize CA transcripts (Burnell and Ludwig, 1997), and has demonstrated the presence of at least three CA genes in the maize genome. Surprisingly, the expression of a single repeat region resulted in an isozyme displaying CA enzymatic activity; in fact bacterial expression of Repeat B produced a protein that exhibited very high levels of activity. The expression of maize CA Repeat B in the cytosol of mesophyll cells of rice plants may be useful in the quest to introduce a C_4 photosynthetic pathway into rice and other agronomically important cereal plants.

References

Alber, B.E., Colangelo, C.M., Dong, J., Stalhandske, C.M.V., Baird, T.T., Tu, C., Fierke, C.A., Silverman, D.N., Scott, R.A. and Ferry, J.G. (1999). *Kinetic and Spectroscopic Characterization of the Gamma-Carbonic Anhydrase from the Methanoarchaeon* Methanosarcina thermophila. Biochemistry, **38**: 13119-13128.

Arai, M., Suzuki, S., Murai, N., Yamada, S., Ohta, S. and Burnell, J.N. (1998). C_4 Cycle of PCK Type. Canada, Japan Tobacco Inc.

Aravind, P. and Prasad, M.N.V. (2004). *Carbonic anhydrase impairment in cadmium-treated* Ceratophyllum demersum L. *(free floating freshwater macrophyte): toxicity reversal by zinc.* Journal of Analytical Atomic Spectrometry, **19**: 52-57.

Aravind, P. and Prasad, M.N.V. (2005). *Zinc mediated protection to the conformation of carbonic anhydrase in cadmium exposed* Ceratophyllum demersum L. Plant Science, **169**: 245-254.

Aronsson, G., Martensson, L.-G., Carlsson, U. and Jonsson, B.-H. (1995). *Folding and Stability of the N-Terminus of Human Carbonic Anhydrase II*. Biochemistry, **34**: 2153-2162.

Ashikawa, I. (2001). *Gene-associated CpG islands in plants as revealed by analyses of genomic sequences.* The Plant Journal, **26**: 617-625.

Atkins, C.A., Patterson, B.D. and Graham, D. (1972a). *Plant Carbonic Anhydrases. 1. Distribution of Types Among Species*. Plant Physiology, **50**: 214-217.

Atkins, C.A., Patterson, B.D. and Graham, D. (1972b). *Plant Carbonic Anhydrases. 2. Preparation and Some Properties of Monocotyledon and Dicotyledon Enzyme Types*. Plant Physiology, **50**: 218-223.

Bailey-Serres, J. and Dawe, R.K. (1996). *Both 5' and 3' Sequences of Maize* adh7 *mRNA Are Required for Enhanced Translation under Low-Oxygen Conditions.* Plant Physiology, **112**: 685-695.

Bar-Akiva, A., Gotfried, A. and Lavon, R. (1971). *A comparison of various means of testing the effectiveness of foliar sprays for correcting zinc deficiencies in citrus trees*. Journal of Horticultural Science, **46**: 397-401.

Benhamed, M., Bertrand, C., Servet, C. and Zhou, D. (2006). Arabidopsis GCN5, HD1, *and* TAF1/HAF2 *Interact to Regulate Histone Acetylation Required for Light-Responsive Gene Expression.* The Plant Cell, **18**: 2893-2903.

Bergenhem, N. (1996). *Chromatographic and electrophoretic methods related to the carbonic anhydrase isozymes*. Journal of Chromatography, **684**: 289-305.

Björkbacka, H., Johansson, I.-M. and Forsman, C. (1999). *Possible Roles for His 208 in the Active-Site Region of Chloroplast Carbonic Anhydrase from Pisum sativum.* Archives of Biochemistry and Biophysics, **361**: 17-24.

Björkbacka, H., Johansson, I.-M., Skärfstad, E. and Forsman, C. (1997). *The Sulfhydryl Groups of Cys 269 and Cys 272 Are Critical for the Oligomeric State of Chloroplastic Carbonic Anhydrase from* Pisum sativum. Biochemistry, **36**: 4287-4294.

Bortiri, E., Jackson, D. and Hake, S. (2006). *Advances in maize genomics: the emergence of positional cloning.* Current Opinion in Plant Biology, **9**: 164-171.

Bracey, M.H., Christiansen, J., Tovar, P., Cramer, S.P. and Bartlett, S.G. (1994). *Spinach Carbonic Anhydrase: Investigation of the Zinc-Binding Ligands by Site-Directed Mutagenesis, Elemental Analysis, and EXAFS.* Biochemistry, **33**: 13126-13131.

Bradford, M.M. (1976). *A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.* Analytical Biochemistry, **7**: 248-254.

Buchanan, C.D., Klein, P.E. and Mullet, J.E. (2004). *Phylogenetic Analysis of 5'-Noncoding Regions From the ABA-Responsive* rab16/17 *Gene Family of Sorghum, Maize and Rice Provides Insight Into the Composition, Organization and Function of* cis-*Regulatory Molecules.* Genetics, **168**: 1639-1654.

Burnell, J.N. (1990). *Immunological Study of Carbonic Anhydrase in C_3 and C_4 Plants Using Antibodies to Maize Cytosolic and Spinach Chloroplastic Carbonic Anhydrase.* Plant Cell Physiology, **31**: 423-427.

Burnell, J.N. (2000). Carbonic Anhydrases of higher plants: an overview. The Carbonic Anhydrases, New Horizons. Chegwidden, W.R., Carter, N.D., and Edwards, Y.H. Basel, Birkhauser Verlag**:** 501-518.

Burnell, J.N., Gibbs, M.J. and Mason, J.G. (1990a). *Spinach Chloroplastic Carbonic Anhydrase - Nucleotide Sequence Analysis of cDNA*. Plant Physiology, **92**: 37-40.

Burnell, J.N., Suzuki, I. and Sugiyama, T. (1990b). *Light Induction and the Effect of Nitrogen Status upon the Activity of Carbonic Anhydrase in Maize Leaves.* Plant Physiology, **94**: 384-387.

Burnell, J.N. and Hatch, M.D. (1988). *Low Bundle Sheath Carbonic Anhydrase is Apparently Essential for Effective C_4 Pathway Operation*. Plant Physiology, **86**: 1252-1256.

Burnell, J.N. and Ludwig, M. (1997). *Characterization of Two cDNAs Encoding Carbonic Anhydrase in Maize Leaves*. Australian Journal of Plant Physiology, **24**: 451-458.

Cavallaro, A., Ludwig, M. and Burnell, J.N. (1994). *The nucleotide sequence of a complementary DNA encoding Flaveria bidentis carbonic anhydrase*. FEBS Letters, **350**: 216-218.

Chen, W-H., Lv, G., Lv, C., Zeng, C. And Hu, S. (2007). *Systematic analysis of alternative first exons in plant genomes.* BMC Plant Biology, **7**: 55-68.

Chen, Z-L., Schuler, M.A. and Beachy, R.N. (1986). *Functional analysis of regulatory elements in a plant embryo-specific gene.* Proceedings of the National Academy of Science, **83**: 8560-8564.

Chen. W. and Singh, K.B. (1999). *The auxin, hydrogen peroxide and salicylic acid induced expression of the Arabidopsis* GST6 *promoter is mediated in part by an ocs element.* The Plant Journal, **19**: 667-677.

Chollet, R., Vidal, J. and O'Leary, M.H. (1996). *PHOSPHOENOLPYRUVATE CARBOXYLASE: A Ubiquitous, Highly Regulated Enzyme in Plants.* Annual Review of Plant Physiology and Plant Molecular Biology, **47**: 273-298.

Chung, H-J., Fu, H-Y. And Thomas, T.L. (2005). *Abscisic acid-inducible nuclear proteins bind to bipartite promoter elements required for ABA response and embryo-regulated expression of the carrot* Dc3 *gene.* Planta, **220**: 424-433.

Covarrubias, A.S., Bergfors, T., Jones, T.A. and Högbom, M. (2006). *Structural Mechanics of the pH-dependent Activity of β-Carbonic Anhydrase from* Mycobacterium tuberculosis. The Journal of Biological Chemistry, **281**: 4993-4999.

Cronk, J.D., Endrizzi, J.A., Cronk, M.R., O'Neill, J.W. and Zhang, K.Y.J. (2001). *Crystal structure of E.coli beta-carbonic anhydrase, an enzyme with an unusual pH-dependent activity.* Protein Science, **10**: 911-922.

Devine, S.E., Chissoe, S.L., Eby, Y., Wilson, R.K. and Boeke, J.D. (1997). *A Transposon-Based Strategy for Sequencing Repetitive DNA in Eukaryotic Genomes.* Genome Research, **7**: 551-563.

Di Fiore, A., Monti, S.M., Hilvo, M., Parkkila, S., Romano, V., Scaloni, A., Pedone, C., Scozzafava, A., Supuran, C.T. and De Simone, G. (2008). *Crystal structure of human carbonic anhydrase XIII and its complex with the inhibitor acetazolamide.* Proteins, **74**: 164-175.

Dolferus, R., Jacobs, M., Peacock, W.J. and Dennis, E.S. (1994). *Differential Interactions of Promoter Elements in Stress Responses of the* Arabidopsis Adh *Gene.* Plant Physiology, **105**: 1075-1087.

Edwards, G.E., Franceschi, V.R. and Voznesenskaya, E.V. (2003). *Single-Cell C_4 Photosynthesis Versus the Duel-Cell (Kranz) Paradigm.* Annual Review of Plant Biology, **55:** 176-196.

Edwards, G.E. and Mohamed, A.K. (1973). *Reduction in Carbonic Anhydrase Activity in Zinc Deficient Leaves of* Phaseolus vulgaris *L.* Crop Science, **13**: 351-354.

Edwards, G.E., Nakamoto, H., Burnell, J.N. and Hatch, M.D. (1985). *Pyruvate,P_i dikinase and NADP-malate dehydrogenase in C_4 photosynthesis: Properties and Mechanism of Light/Dark Regulation.* Annual Review of Plant Biology, **36**: 255-286.

Edwards, J.W., Walker, E.L. and Coruzzi, G.M. (1990). *Cell-specific expression in transgenic plants reveals nonoverlapping roles for chloroplast and cytosolic glutamine synthetase.* Proceedings of the National Academy of Science, **87**: 3459-3463.

Elleby, B., Chirica, L.C., Tu, C., Zeppezauer, M. and Lindskog, S. (2001). *Characterization of carbonic anhydrase from* Neisseria gonorrhoeae. European Journal of Biochemistry, **268**: 1613-1619.

Emrich, S.J., Li, L., Wen, T-J., Yandeau-Nelson, M.D., Fu, Y., Guo, L., Chou, H-H., Aluru, S., Ashlock, D.A. and Schnable, P.S. (2007). *Nearly Identical Paralogs: Implications for Maize (*Zea mays *L.) Genome Evolution.* Genetics, **175**: 429-439.

Eriksson, M., Villand, P., Gardestrom, P. and Samuelsson, G. (1998). *Induction and Regulation of Expression of a Low-CO_2-Induced Mitochondrial Carbonic Anhydrase in Chlamydomonas reinhardtii*. Plant Physiology, **116**: 637-641.

Everson, R.G. and Slack, C.R. (1968). *Distribution of a Carbonic Anhydrase in Relation to the C_4 Pathway of Photosynthesis*. Phytochemistry, **7**: 581-584.

Ewing, R.M., Jenkins, G.I. and Langdale, J.A. (1998). *Transcripts of maize* RbcS *genes accumulate differentially in C_3 and C_4 tissues.* Plant Molecular Biology, **36**: 593-599.

Fabre, N., Reiter, I.M., Becuwe-Linka, N., Genty, B. and Rumeau, D. (2007). *Characterization and expression analysis of genes encoding α and β carbonic anhydrases in* Arabidopsis. Plant, Cell and Environment, **30**: 617-629.

Fawcett, T.W., Browse, J.A., Volokita, M. and Bartlett, S.G. (1990). *Spinach Carbonic Anhydrase Primary Structure Deduced from the Sequence of a cDNA Clone*. The Journal of Biological Chemistry, **265**: 5414-5417.

Fernley, R.T. (1988). *Non-cytoplasmic carbonic anhydrases*. Trends in Biochemical Science, **13**: 326-359.

Fett, J.P. and Coleman, J.R. (1994). *Characterization and Expression of Two cDNAs Encoding Carbonic Anhydrase in* Arabidopsis thaliana. Plant Physiology, **105**: 707-713.

Fisher, S.Z., Tariku, I., Case, N.M., Tu, C., Seron, T., Silverman, D., Linser, P.J. and McKenna, R. (2006). *Expression, purification, kinetic, and structural characterization of an α-class carbonic anhydrase from* Aedes aegypti (*AaCA1*). Biochimica et Biophysica Acta, **1764**: 1413-1419.

Friso, G., Giacomelli, L., Ytterberg, A.J., Peltier, J-B., Rudella, A., Sun, Q. and van Wilk, K.J. (2004). *In-Depth Analysis of the Thylakoid Membrane Proteome of* Arabidopsis thaliana *Chloroplasts: New Proteins, New Functions, and a Plastid Proteome Database.* The Plant Cell, **16**: 478-499.

Furumoto, T., Hata, S. and Izui, K. (1999). *cDNA cloning and characterization of maize phosphoenolpyruvate carboxykinase, a bundle sheath cell-specific enzyme.* Plant Molecular Biology, **41**: 301-311.

Gàlvez, S., Hirsch, A.M., Wycoff, K.L., Hunt, S., Layzell, D.B., Kondorosi, A. and Crespi, M. (2000). *Oxygen Regulation of a Nodule-Located Carbonic Anhydrase in Alfalfa*. American Society of Plant Physiologists, **124**: 1059-1068.

Gardiner, J., Schroeder, S., Polacco, M.L., Sanchez-Villeda, H., Fang, Z., Morgante, M., Landewe, T., Fengler, K., Useche, F., Hanafey, M., Tingey, S., Chou, H., Wing, R., Soderlund, C. and Coe, E.H. Jr. (2004). *Anchoring 9,371 Maize Expressed Sequence Tagged Unigenes to the Bacterial Artificial Chromosome Contig Map by Two-Dimensional Overgo Hybridization.* Plant Physiology, **134**: 1317-1326.

Gaut, B.S. and Doebley, J.F. (1997). *DNA sequence evidence for the segmental allotetraploid origin of maize.* Proceedings of the National Academy of Science, **94**: 6809-6814.

Giordano, M., Norici, A., Forssen, M., Eriksson, M. and Raven, J. (2003). *An anaplerotic role for mitochondrial carbonic anhydrase in Chlamydomonas reinhardtii*. Plant Physiology, **132**: 2126-2134.

Golovkin, M. and Reddyl, A.S.N. (1996). *Structure and Expression of a Plant U1 snRNP 70K Gene: Alternative Splicing of U1 snRNP 70K Pre-mRNAs Produces Two Different Transcripts.* The Plant Cell, **8**: 1421-1435.

Gómez-Maldonado, J., Avila, C., de la Torre, F., Caňas, R., Cánovas, F.M. and Campbell, M.M. (2004). *Functional interactions between a glutamine synthetase promoter and MYB proteins.* The Plant Journal, **39**: 513-526.

Goodall, G.J. and Filipowicz, W. (1991). *Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants.* The EMBO Journal, **10**: 2635-2644.

Graham, D., Reed, M.L., Patterson, B.D. and Hockley, D.G. (1984). *Chemical Properties, Distribution and Physiology of Plant and Algal Carbonic Anhydrases*. Annals of the New York Academy of Science, **429**: 222-237.

Guliev, N.M., Babaev, G.G., Bairamov, S.M. and Aliev, D.A. (2003). *Purification, Properties, and Localization of Two Carbonic Anhydrases from* Amaranthus cruentus *Leaves*. Russian Journal of Plant Physiology, **50**: 213-219.

Guo, M., Rupe, M.A., Zinselmeier, C., Habben, J., Bowen, B.A. and Smith, O.S. (2004). *Allelic Variation of Gene Expression in Maize Hybrids.* The Plant Cell, **16**: 1707-1716.

Hall, R.A., Vullo, D., Innocenti, A., Scozzafava, A., Supuran, C.T., Klappa, P. and Mühlschlegel, F.A. (2008). *External pH influences the transcriptional profile of the carbonic anhydrase, CAH-4b in* Caenorhabditis elegans. Molecular and Biochemical Parasitology, **161**: 140-149.

Hanley, B.A. and Schuler, M.A. (1988). *Plant intron sequences: evidence for distinct groups of introns.* Nucleic Acids Research, **16**: 7159-7176.

Hansen, E.R., Petracek, M.E., Dickey, L.F. and Thompson, W.F. (2001). *The 5' End of the Pea Ferredoxin-1 mRNA Mediates Rapid and Reversible Light-Directed Changes in Translation in Tobacco.* Plant Physiology, **125**: 770-778.

Hanson, D.T., Franklin, L.A., Samuelsson, G. and Badger, M.R. (2003). *The* Chlamydomonas reinhardtii *cia3 Mutant Lacking a Thylakoid Lumen-Localized Carbonic Anhydrase Is Limited by CO_2 Supply to Rubisco and Not Photosystem II Function* in vivo. Plant Physiology, **132**: 2267-2275.

Harada, H., Nakatsuma, D., Ishida, M. and Matsuda, Y. (2005). *Regulation of the Expression of Intracellular β-Carbonic Anhydrase in Reponse to CO_2 and Light in the Marine Diatom* Phaeodactylum tricornutum. Plant Physiology, **139**: 1041-1050.

Harter, K., Kircher, S., Frohnmeyer, H., Krenz, M., Nagy, F. and Schafer, E. (1994). *Light-Regulated Modification and Nuclear Translocation of Cytosolic G-Box Binding Factors in Parsley.* The Plant Cell, **6**: 545-559.

Hartmann, U., Sagasser, M., Mehrtens, F., Stracke, R. and Weisshaar, B. (2005). *Differential combinatorial interactions of* cis-*acting elements recognized by R2R3-MYB, BZIP, and BHLH factors control light-responsive and tissue-specific activation of phenylpropanoid biosynthesis genes.* Plant Molecular Biology, **57**: 155-171.

Hatch, M.D. (1991). *Carbonic Anhydrase Assay: Strong Inhibition of the Leaf Enzyme by CO_2 in Certain Buffers.* Analytical Biochemistry, **192**: 85-89.

Hatch, M.D. and Burnell, J.N. (1990). *Carbonic Anhydrase in Leaves and Its Role in the First Step of C_4 Photosynthesis.* Plant Physiology, **93**: 825-828.

Hatch, M.D., Slack, C.R. and Johnson, H.S. (1967). *Further Studies on a New Pathway of Photosynthetic Carbon Dioxide Fixation in Sugar-Cane and its Occurrence in other Plant Species*. Biochemistry Journal, **102**: 417-422.

Hausler, R.E., Holtum, J.A.M. and Latzko, E. (1987). *CO_2 is the inorganic carbon substrate of NADP malic enzymes from* Zea mays *and from wheat germ*. European Journal of Biochemistry, **163**: 619-626.

Hewett-Emmett, D. and Tashian, R.E. (1996). *Function Diversity, Conservation, and Convergence in the Evolution of the α-, β-, and γ-Carbonic Anhydrase Gene Families*. Molecular Phylogenetics and Evolution, **5**: 50-77.

Hiltonen, T., Björkbacka, H., Forsman, C., Clarke, A.K. and Samuelsson, G. (1998). *Intracellular β-Carbonic Anhydrase of the Unicellular Green Alga* Coccomyxa. Plant Physiology, **117**: 1341-1349.

Hilvo, M., Baranauskiene, L., Salzano, A.M., Scaloni, A., Matulis, D., Innocenti, A., Scozzafava, A., Monti, S.M., Di Fiore, A., De Simone, G., Lindfors, M., Jänis, J., Valjakka, J., Pastoreková, S., Pastorek, J., Kulomaa, M.S., Nordlund, H.R., Supuran, C.T. and Parkkila, S. (2008). *Biochemical Characterization of CA IX, One of the Most Active Carbonic Anhydrase Isozymes.* The Journal of Biological Chemistry, **283**: 27799-27809.

Hoang, C.V. and Chapman, K.D. (2002). *Regulation of carbonic anhydrase gene expression in cotyledons of cotton (*Gossypium hirsutum *L.) seedlings during post-germinative growth.* Plant Molecular Biology, **49**: 449-458.

Hoang, C.V., Wessler, H.G., Local, A., Turley, R.B., Benjamin, R.C. and Chapman, K.D. (1999). *Identification and Expression of Cotton (*Gossypium hirsutum *L.) Plastidial Carbonic Anhydrase*. Plant Cell Physiology, **40**: 1262-1270.

International Rice Genome Sequencing Project. (2005). *The map-based sequence of the rice genome.* Nature, **436**: 793-800.

Iverson, T.M., Alber, B.E., Kisker, C., Ferry, J.G. and Rees, D.C. (2000). *A Closer Look at the Active Site of γ-Class Carbonic Anhydrases: High-Resolution Crystallographic Studies of the Carbonic Anhydrase from* Methanosarcina thermophila. Biochemistry, **39**: 9222-9231.

Jenkins, C.L., Burnell, J.N. and Hatch, M.D. (1987). *Form of Inorganic Carbon Involved as a Product and as an Inhibitor of C_4 Acid Decarboxylases Operating in C_4 Photosynthesis*. Plant Physiology, **85**: 952-957.

Jiang, W. and Gupta, D. (1999). *Structure of the carbonic anhydrase VI (CA6) gene: evidence for two distinct groups within the α-CA gene family.* Biochemical Journal, **344**: 385-390.

Karlsson, J., Clarke, A.K., Chen, Z.-Y., Hugghins, S.Y., Park, Y.-I., Husic, H.D., Moroney, J.V. and Samuelsson, G. (1998). *A novel α-type carbonic anhydrase associated with the thylakoid membrane in* Chlamydomonas reinhardtii *is required for growth at ambient CO_2.* The EMBO Journal, **17**: 1208-1216.

Karlsson, J., Hiltonen, T., Husic, D., Ramazanov, Z. and Samuelsson, G. (1995). *Intracellular Carbonic Anhydrase of* Chlamydomonas reinhardtii. Plant Physiology, **109**: 533-539.

Kavroulakis, N., Flemetakis, E., Aivalakis, G. and Katinakis, P. (2000). *Carbon Metabolism in Developing Soybean Root Nodules: The Role of Carbonic Anhydrase.* Molecular Plant-Microbe Interactions, **13**: 14-22.

Kebeish, R., Niessen, M., Thiruveedhi, K., Bari, R., Hirsch, H.J., Rosenkranz, R., Stäbler, N., Schönfeld, B., Kreuzaler, F. and Peterhänsel, C. (2007). *Chloroplastic photorespiratory bypass increases photosynthesis and biomass production in* Arabidopsis thaliana. Nature Biotechnology, **25**: 539-540.

Khan, M.S. (2007). *Engineering photorespiration in chloroplasts: a novel strategy for increasing biomass production.* Trends in Biotechnology, **25**: 437-440.

Kimber, M.S. and Pai, E.F. (2000). *The active site architecture of* Pisum sativum *β-carbonic anhydrase is a mirror image of that of α-carbonic anhydrase*. The EMBO Journal, **19**: 1407-1418.

Kisiel, W. and Graf, G. (1972). *Purification and Characterisation of Carbonic Anhydrase from* Pisum sativum*.* Phytochemistry, **11**: 113-117.

Kucho, K., Yoshioka, S., Taniguchi, F., Ohyama, K. and Fukuzawa, H. (2003). *Cis-acting Elements and DNA-Binding Proteins Involved in CO_2-Responsive Transcriptional Activation of* Cah1 *Encoding a Periplasmic Carbonic Anhydrase in* Chlamydomonas reinhardtii. Plant Physiology, **133**: 783-793.

Kusian, B., Sultemeyer, D. and Bowien, B. (2002). *Carbonic Anhydrase Is Essential for Growth of* Ralstonia eutropha *at Ambient CO_2 Concentrations.* Journal of Bacteriology, **184**: 5018-5026.

Kyozuka, J., McElroy, D., Hayakawa, T., Xie, Y., Wu, R. and Shimamoto, K. (1993). *Light-Regulated and Cell-Specific Expression of Tomato* rbcS-gusA *and Rice* rbcS-gusA *Fusion Genes in Transgenic Rice.* Plant Physiology, **102**: 991-1000.

Kyozuka, J., Olive, M., Peacock, J., Dennis, E.S. and Shimamoto, K. (1994). *Promoter Elements Required for Developmental Expression of the Maize* Adh1 *Gene in Transgenic Rice.* The Plant Cell, **6**:799-810.

Kyte, J. and Doolittle, R.F. (1982). *A simple method for displaying the hydropathic character of a protein.* Journal of Molecular Biology, **157**: 105-132.

Lal, S.K. and Hannah, L.C. (2005). *Helitrons contribute to the lack of gene colinearity observed in modern maize inbreds.* Proceedings of the National Academy of Science, **102**: 9993-9994.

Lane, T.W., Saito, M.A., George, G.N., Pickering, I.J., Prince, R.C. and Morel, F.M. (2005). *Biochemistry: a cadmium enzyme from a marine diatom.* Nature, **435**: 42.

Langdale, J.A., Zelitch, I., Miller, E. and Nelson, T. (1988a). *Cell position and light influence C_4 versus C_3 patterns of photosynthetic gene expression in maize.* The EMBO Journal, **7**: 3643-3651.

Langdale, J.A., Rothermel, B.A. and Nelson, T. (1988b). *Cellular pattern of photosynthetic gene expression in developing maize leaves.* Genes and Development, **2**: 106-115.

Lapointe, M., Mackenzie, T.D. and Morse, D. (2008). *An external δ-carbonic anhydrase in a free-living marine dinoflagellate may circumvent diffusion-limited carbon acquisition.* Plant Physiology, **147**: 1427-1436.

Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992). *CpG islands as gene markers in the human genome.* Genomics, **13**: 1095-1107.

Lawlor, D.W. (2001). Photosynthesis. London, Bios Scientific Publishers.

Lazova, G.N. and Stemler, A.J. (2008). *A 160 kDa protein with carbonic anhydrase activity is complexed with Rubisco on the outer surface of thylakoids.* Cell Biology International, **32**: 646-653.

Lea, P.J., Chen, Z.-H., Leegood, R.C. and Walker, R.P. (2001). *Does phosphoenolpyruvate carboxykinase have a role in both amino acid and carbohydrate metabolism?* Amino Acids, **20**: 225-241.

Lea, P.J. and Leegood, R.C., Eds. (1999). Plant Biochemistry and Molecular Biology. Chichester, Wiley.

Lebrun, M., Waksman, G. and Freyssinet, G. (1987). *Nucleotide sequence of a gene encoding corn ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit (rbcs).* Nucleic Acids Research, **15**: 4360.

Lees, J.G. and Janes, R.W. (2008). *Combining sequence-based prediction methods and circular dichroism and infrared spectroscopic data to improve protein secondary structure determinations.* BMC Bioinformatics, **9**: 24-30.

Leggat, W., Dixon, R., Saleh, S. and Yellowlees, D. (2005). *A novel carbonic anhydrase from the giant clam* Tridacna gigas *contains two carbonic anhydrase domains.* FEBS Journal, **272**: 3297-3305.

Lehtonen, J., Shen, B., Vihinen, M., Casini, A., Scozzafava, A., Supuran, C.T., Parkkila, A.K., Saarnio, J., Kivela, A.J., Waheed, A., Sly, W.S. and Parkkila, S. (2004). *Characterization of CA XIII, a novel member of the carbonic anhydrase isozyme family*. Journal of Biological Chemistry, **279**: 2719-2727.

Lorimer, G.H. (1981). *The carboxylation and oxygenation of ribulose 1,5-bisphosphate: the primary events in photosynthesis and photorespiration.* Annual Review of Plant Physiology, **32**: 349-383.

Lu, Y.-K. and Stemler, A.J. (2002). *Extrinsic Photosystem II Carbonic Anhydrase in Maize Mesophyll Chloroplasts*. Plant Physiology, **128**: 643-649.

Ludwig, M., von Caemmerer, S., Price, D., Badger, M.R. and Furbank, R.T. (1998). *Expression of Tobacco Carbonic Anhydrase in the C_4 Dicot* Flaveria bidentis *Leads to Increased Leakiness of the Bundle Sheath and a Defective CO_2-Concentrating Mechanism*. Plant Physiology, **117**: 1071-1081.

Ma, J., SanMiguel, P.S., Lai, J., Messing, J. and Bennetzen, J.L. (2005). *DNA Rearrangement in Orthologous* Orp *Regions of the Maize, Rice and Sorghum Genomes.* Genetics, **170**: 1209-1220.

Majeau, N. and Coleman, J.R. (1992). *Nucleotide Sequence of a Complementary DNA Encoding Tobacco Chloroplastic Carbonic Anhydrase.* Plant Physiology, **100**: 1077-1078.

Majeau, N., Arnolko, M.A. and Coleman, J.R. (1994). *Modification of carbonic anhydrase activity by antisense and over-expression constructs in transgenic tobacco.* Plant Molecular Biology, **25**: 377-385.

Majeau, N. and Coleman, J.R. (1994). *Correlation of Carbonic Anhydrase and Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase Expression in Pea.* Plant Physiology, **104**: 1393-1399.

Makino, A. and Sage, R.F. (2007). *Temperature Response of Photosynthesis in Transgenic Rice Transformed with 'Sense' or 'Antisense'* rbcS. Plant Cell Physiology, **48**: 1472-1483.

Martineau, B. and Taylor, W.C. (1985). *Photosynthetic Gene Expression and Cellular Differentiation in Developing Maize Leaves.* Plant Physiology, **78**: 399-404.

Martineau, B. and Taylor, W.C. (1986). *Cell-Specific Photosynthetic Gene Expression in Maize Determined Using Cell Separation Techniques and Hybridization* in Situ. Plant Physiology, **82**: 613-618.

Massonneau, A., Bouba-Hérin, N., Pethe, C., Madzak, C., Falque, M., Mercy, M., Kopecny, D., Majira, A., Rogowsky, P. and Laloue, M. (2004). *Maize cytokinin oxidase genes: differential expression and cloning of two new cDNAs.* Journal of Experimental Botany, **55**: 2549-2557.

Mathews, C.K. and Van Holde, K.E. (1996). Biochemistry. New York, The Benjamin/Cummings Publishing Company Inc.

Matos, A.R., d'Arcy-Lameta, A., França, M., Pêtres, S., Edelman, L., Kader, J-C., Zuily-Fodil, Y. and Pham-Thi, A.T. (2001). *A novel patatin-like gene stimulated by drought stress encodes a galactolipid acyl hydrolase.* FEBS Letters, **491**: 188-192.

Matsuoka, M. (1990). *Structure, Genetic Mapping, and Expression of the Gene for Pyruvate, Orthophosphate Dikinase from Maize.* The Journal of Biological Chemistry, **265**: 16772-16777.

McCall, K.A., Huang, C.-C. and Fierke, C.A. (2000). *Function and Mechanism of Zinc Metalloenzymes*. American Society for Nutrition and Health, **Supplement**: 1437S - 1446S.

McCullough, A.J., Baynton, C.E. and Shuler, M.A. (1996). *Interactions across Exons Can Influence Splice Site Recognition in Plant Nuclei.* The Plant Cell, **8**: 2295-2307.

McGinn, P.J. and Morel, F.M.M. (2008). *Expression and regulation of carbonic anhydrases in the marine diatom* Thalassiosira pseudonana *and in natural phytoplankton assemblages from Great Bay, New Jersey.* Physiologia Plantarum, **133**: 78-91.

Meldrum, N.U. and Roughton, F.J.W. (1933). *Carbonic Anhydrase. Its Preparation and Properties.* Journal of Physiology, 113-142.

Messing, J., Bharti, A.K., Karlowski, W.M., Gundlach, H., Kim, H.R., Yu, Y., Wei, F., Fuks, G., Soderlund, C.A., Mayer, K.F.X. and Wing, R.A. (2004). *Sequence composition and genome organization of maize.* Proceedings of the National Academy of Science, **101**: 14349-14354.

Miflin, B.J. and Habash, D.Z. (2002). *The role of glutamine synthetase and glutamate dehydrogenase in nitrogen assimilation and possibilities for improvement in the nitrogen utilization of crops.* Journal of Experimental Botany, **53**: 979-987.

Mitra, M., Lato, S.M., Ynalvez, R.A., Xiao, X. and Moroney, J.V. (2004). *Identification of a New Chloroplastic Carbonic Anhydrase in* Chlamydomonas reinhardtii. Plant Physiology, **135**: 173-182.

Mitsuhashi, S. and Miyachi, S. (1996). *Amino Acid Sequence Homology between N- and C-terminal Halves of a Carbonic Anhydrase in* Porphyridium purpureum*, as Deduced from the Cloned cDNA.* The Journal of Biological Chemistry, **271**: 28703-28709.

Mitsuhashi, S., Mizushima, T., Yamashita, E., Yamamoto, M., Kumasaka, T., Moriyama, H., Ueki, T., Miyachi, S. and Tsukihara, T. (2000). *X-ray Structure of β-Carbonic Anhydrase from the Red Alga,* Porphyridium purpureum*, Reveals a Novel Catalytic Site for CO_2 Hydration.* The Journal of Biological Chemistry, **275**: 5521-5526.

Miyao, M. (2003). *Molecular evolution and genetic engineering of C_4 photosynthetic enzymes*. Journal of Experimental Botany, **54**: 179-189.

Mohanty, B., Krishnan, S.P.T., Swarup, S. and Bajic, V.B. (2005). *Detection and Preliminary Analysis of Motifs in Promoters of Anaerobically Induced Genes of Different Plant Species.* Annals of Botany, **96**: 669-681.

Morishige, D.T., Childs, K.L., Moore, D. and Mullet, J.E. (2002). *Targeted Analysis of Orthologous* Phytochrome A *Regions of the Sorghum, Maize and Rice Genomes using Comparative Gene-Island Sequencing.* Plant Physiology, **130**: 1614-1625.

Moroney, J.V., Bartlett, S.G. and Samuelsson, G. (2001). *Carbonic anhydrases in plants and algae*. Plant, Cell and Environment, **24**: 141-159.

Moubarak-Milad, M. and Stemler, A. (1994). *Oxidation-Reduction Potential Dependence of Photosystem II Carbonic Anhydrase in Maize Thylakoids.* Biochemistry, **33**: 4432-4438.

Moya, A., Tambutté, S., Bertucci, A., Tambutté, E., Lotto, S., Vullo, D., Supuran, C.T., Allemand, D. and Zoccola, D. (2008). *Carbonic anhydrase in the scleractinian coral* Stylophora pistillata: *characterization, localization, and role in biomineralization.* Journal of Biological Chemistry, **283**: 25475-25484.

Mukherjee, K., Choudhury, A.R., Gupta, B., Gupta, S. and Sengupta, D.N. (2006). *An ABRE-binding factor, OSBZ8, is highly expressed in salt tolerant cultivars than in salt sensitive cultivars of indica rice.* BMC Plant Biology, **6**: 18-32.

Ohki, K. (1976). *Effect of Zinc Nutrition on Photosynthesis and Carbonic Anhydrase Activity in Cotton*. Plant Physiology, **38**: 300-304.

Okabe, K., Yang, S.-Y., Tsuzuki, M. and Miyachi, S. (1984). *Carbonic Anhydrase: Its content in spinach leaves and its taxonomic diversity studied with anti-spinach leaf carbonic anhydrase antibody.* Plant Science Letters, **33**: 145-153.

Olive, M.R., Peacock, W.J. and Dennis, E.S. (1991). *The anaerobic responsive element contains two GC-rich sequences essential for binding a nuclear protein and hypoxic activation of the maize* Adh1 *promoter.* Nucleic Acids Research, **19**: 7053-7060.

Parisi, G., Perales, M., Fornasari, M.S., Colaneri, A., Gonzalez-Schain, N., Gomez-Casati, D., Zimmermann, S., Brennicke, A., Araya, A., Ferry, J.G., Echave, J. and Zabaleta, E. (2004). *Gamma carbonic anhydrases in plant mitochondria*. Plant Molecular Biology, **55**: 193-207.

Park, Y.-I., Karlsson, J., Rojdestvenski, I., Pronina, N., Klimov, V., Oquist, G. and Samuelsson, G. (1999). *Role of a novel photosystem II-associated carbonic anhydrase in photosynthetic carbon assimilation in Chlamydomonas reinhardtii.* FEBS Letters, **444**: 102-105.

Patel, M., Slegel, A.J. and Berry, J.O. (2006). *Untranslated Regions of* FbRbcS1 *mRNA Mediate Bundle Sheath Cell-specific Gene Expression in Leaves of a C_4 Plant.* Journal of Biological Chemistry, **281**: 25485-25491.

Perales, M., Eubel, H., Heinemeyer, J., Colaneri, A., Zabaleta, E. and Braun, H-P. (2005). *Disruption of a Nuclear Gene Encoding a Mitochondrial Gamma Carbonic Anhydrase Reduces Complex I and Supercomplex I + III2 Levels and Alters Mitochondrial Physiology in Arabidopsis.* Journal of Molecular Biology, **350**: 263-277.

Peters, K., Dudkina, N.V. Jänsch, L., Braun, H.P. and Boekema, E.J. (2008). *A structural investigation of complex I and I+III2 supercomplex from* Zea mays *at 11 – 13 Å resolution: assignment of the carbonic anhydrase domain and evidence for structural heterogeneity within complex I.* Biochimica et Biophysica Acta, **1777**: 84-93.

Pocker, Y. and Ng, J.S.Y. (1973). *Plant Carbonic Anhydrase. Properties and Carbon Dioxide Hydration Kinetics.* Biochemistry, **12**: 5127-5134.

Poincelot, R.P. (1972). *The Distribution of Carbonic Anhydrase and Ribulose Diphosphate Carboxylase in Maize Leaves.* Plant Physiology, **59**: 336-340.

Reed, M.L. and Graham, D. (1980). *Carbonic Anhydrase in plants: distribution, properties and possible physiological roles*. Progress in Phytochemistry, **7**: 47-94.

Roeske, C.A. and Ogren, W.L. (1990). *Nucleotide sequence of pea cDNA encoding chloroplast carbonic anhydrase*. Nuclei Acids Research, **18**: 3413.

Rook, F., Hadingham, S.A., Li, Y. and Bevan, M.W. (2006). *Sugar and ABA responsive pathways and the control of gene expression.* Plant, Cell and Environment, **29**: 426-434.

Rost, B. (1996). *PHD: predicting one-dimensional protein structure by profile-based neural networks.* Methods in Enzymology, **266**: 525-539.

Rothermel, B.A. and Nelson, T. (1989). *Primary Structure of the Maize NADP-dependent Malic Enzyme*. The Journal of Biological Chemistry, **264**: 19587-19592.

Rouhier, N., Villarejo, A., Srivastava, M., Gelhaye, E., Keech, O., Droux, M., Finkemeier, I., Samuelsson, G., Dietz, K.J., Jacquot, J-P. and Wingsle, G. (2005). *Identification of Plant Glutaredoxin Targets.* Antioxidants and Redox Signalling, **7**: 919-929.

Rowlett, R.S., Chance, M.R., Wirt, M.D., Sidelinger, D.E., Royal, J.R., Woodroffe, M., Wang, Y.-F.A., Saha, R.P. and Lam, M.G. (1994). *Kinetic and Structural Characterisation of Spinach Carbonic Anhydrase*. Biochemistry, **33**: 13967-13976.

Rowlett, R.S., Tu, C., McKay, M.M., Preiss, J.R., Loomis, R.J., Hicks, K.A., Marchione, R.J., Strong, J.A., Donovan, G.S.J. and Chamberlin, J.E. (2002). *Kinetic characterization of wild-type and proton transfer-impaired variants of β-carbonic anhydrase from* Arabidopsis thaliana*.* Archives of Biochemistry and Biophysics, **404**: 197-209.

Rudenko, N.N., Ignatova, L.K. and Ivanov, B.N. (2007). *Multiple sources of carbonic anhydrase activity in pea thylakoids: soluble and membrane-bound forms.* Photosynthesis Research, **91**: 81-89.

Sage, R.F. (2004). *The evolution of C_4 photosynthesis*. New Phytologist, **161**: 341-370.

Sage, R.F. and Monson, R.K. (1999). C_4 Plant Biology. San Diego, Academic Press.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual. 2nd ed. New York, Cold Spring Harbor Laboratory Press.

SanMiguel, P. and Bennetzen, J.L. (1998). *Evidence that a Recent Increase in Maize Genome Size was Caused by the Massive Amplification of Intergene Retrotransposons.* Annals of Botany, **82**: 37-44.

Satoh, D., Hiraoka, Y., Colman, B. and Matsuda, Y. (2001). *Physiological and Molecular Biological Characterization of Intracellular Carbonic Anhydrase from the Marine Diatom* Phaeodactylum tricornutum*.* Plant Physiology, **126**: 1459-1470.

Seki, H., Nagasugi, Y., Ichinose, Y., Shiraishi, T. and Yamada, T. (1999). *Changes in* in vivo *DNA-protein Interactions in Pea Phenylalanine Ammonia-lyase and Chalcone Synthase Gene Promoter Induced by Fungal Signal Molecules.* Plant Cell Physiology, **40**: 88-95.

Shahmuradov, I.A., Gammerman, A.J., Hancock, J.M., Bramley, P.M. and Solovyev, V.V. (2003). *PlantProm: a database of plant promoter gene sequences.* Nucleic Acids Research, **31**: 114-117.

Sheehy, J.E., Mitchell, P.L. and Hardy, B. (Eds). (2007). Charting new pathways to C_4 rice. Los Baños (Philippines): International Rice Research Institute.

Sheen, J. (1991). *Molecular Mechanisms Underlying the Differential Expression of Maize Pyruvate, Orthophosphate Dikinase Genes.* The Plant Cell, **3**: 225-245.

Sheen, J.-Y. and Bogorad, L. (1987). *Differential Expression of C_4 Pathway Genes in Mesophyll and Bundle Sheath Cells of Greening Maize Leaves*. The Journal of Biological Chemistry, **262**: 11726-11730.

Silverman, D.N. (1991). *The catalytic mechanism of carbonic anhydrase*. Canadian Journal of Botany, **69**: 1070-1078.

Sjöblom, B., Elleby, B., Wallgren, K., Jonsson, B-H. And Lindskog, S. (1996). *Two point mutations convert a catalytically inactive carbonic anhydrase-related protein (CARP) to an active enzyme.* FEBS Letters, **398**: 322-325.

Slaymaker, D.H., Navarre, D.A., Clarke, D., del Pozo, O., Martin, G.B. and Klessig, D.F. (2002). *The tobacco salicylic acid-binding protein 3 (SABP3) is the chloroplast carbonic anhydrase, which exhibits antioxidant activity and plays a role in the hypersensitive defense response.* Proceedings of the National Academy of Science, **99**: 11640-11645.

Sly, W.S. and Hu, P.Y. (1995). *Human Carbonic Anhydrases and Carbonic Anhydrase Deficiencies*. Annual Review of Biochemistry, **64**: 375-401.

Song, R. and Messing, J. (2003). *Gene expression of a gene family in maize based on noncollinear haplotypes.* Proceedings of the National Academy of Science, **100**: 9055-9060.

Soto, A.R., Zheng, H., Shoemaker, D., Rodriguez, J., Read, B.A. and Wahlund, T.M. (2006). *Identification and Preliminary Characterization of Two cDNAs Encoding Unique Carbonic Anhydrases from the Marine Alga* Emiliania huxleyi. Applied and Environmental Microbiology, **72**: 5500-5511.

Springer, N.M., Xu, X., and Barbazuk, W.B. (2004). *Utility of Different Gene Enrichment Approaches Toward Identifying and Sequencing the Maize Gene Space.* Plant Physiology, **136**: 3023-3033.

Stemler, A.J. (1997). *The case for chloroplast thylakoid carbonic anhydrase*. Physiologia Plantarum, **99**: 348-353.

Strathmann, M., Hamilton, B.A., Mayeda, C.A., Simon, M.I., Meyerowitz, E.M. and Palazzolo, M.J. (1991). *Transposon-facilitated DNA sequencing.* Proceedings of the National Academy of Science, **88**: 1247-1250.

Sugiharto, B., Burnell, J.N. and Sugiyama, T. (1992a). *Cytokinin is Required to Induce the Nitrogen-Dependent Accumulation of mRNAs for Phosphoenolpyruvate Carboxylase and Carbonic Anhydrase in Detached Maize Leaves.* Plant Physiology, **100**: 153-156.

Sugiharto, B., Suzuki, I., Burnell, J.N. and Sugiyama, T. (1992b). *Glutamine Induces the N-Dependent Accumulation of mRNAs Encoding Phosphoenolpyruvate Carboxylase and Carbonic Anhydrase in Detached Maize Leaf Tissue.* Plant Physiology, **100**: 2066-2070.

Sultemeyer, D., Schmidt, C. and Fock, H.P. (1993). *Carbonic anhydrases in higher plants and aquatic microorganisms.* Physiologia Plantarum, **88**: 179-190.

Sunderhaus, S., Dudkina, N.V., Jänsch, L., Klodmann, J. Heinemeyer, J., Perales, M., Zabaleta, E., Boekema, E.J. and Braun, H-P. (2006). *Carbonic Anhydrase Subunits Form a Matrix-exposed Domain Attached to the Membrane Arm of Mitochondrial Complex I in Plants.* The Journal of Biological Chemistry, **281**: 6482-6488.

Supuran, C.T. (2008). *Carbonic anhydrases: novel therapeutic applications for inhibitors and activators.* Nature Reviews Drug Discovery, **7**: 168-181.

Supuran, C.T., Scozzafava, A. and Casini, A. (2003). *Carbonic Anhydrase Inhibitors.* Medicinal Research Reviews, **23**: 146-189.

Suzuki, S. and Burnell, J.N. (1995). *Nucleotide Sequence of a cDNA Encoding Rice Chloroplastic Carbonic Anhydrase*. Plant Physiology, **107**: 299-300.

Swigonová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L. and Messing, J. (2004). *Close Split of Sorghum and Maize Genome Progenitors.* Genome Research, **14**: 1916-1923.

Taniguchi, Y., Nagasaki, J., Kawasaki, M., Miyake, H., Sugiyama, T. and Taniguchi, M. (2004). *Differentiation of Dicarboxylate Transporters in Mesophyll and Bundle Sheath Chloroplasts of Maize*. Plant Cell Physiology, **45**: 187-200.

Tashian, R.E. (1992). Genetics of the Mammalian Carbonic Anhydrases. Advances in Genetics. London, Academic Press Inc. **30:** 321-356.

Tetu, S.G., Tanz, S.K., Vella, N., Burnell, J.N. and Ludwig, M. (2007). *The* Flaveria bidentis *β-Carbonic anhydrase Gene Family Encodes Cytosolic and Chloroplastic Isoforms Demonstrating Distinct Organ-Specific Expression Patterns.* Plant Physiology, **144**: 1316-1327.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Research, **22**: 4673-4680.

Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y., Gorenstein, N.M., Bennetzen, J.L. and Avramova, Z. (1999). *Colinearity and its exceptions in orthologous* adh *regions of maize and sorghum.* Proceedings of the National Academy of Science, **96**: 7409-7414.

Triolo, L., Bagnara, L., Anselmi, C. and Bassanelli, C. (1974). *Carbonic Anhydrase Activity and Localisation in Some Plant Species*. Physiologia Plantarum, **31**: 86-89.

Tripp, B.C., Bell, C.B., Cruz, F., Krebs, C. and Ferry, J.G. (2004). *A Role for Iron in an Ancient Carbonic Anhydrase.* The Journal of Biological Chemistry, **279**: 6683-6687.

Tripp, B.C., Smith, K. and Ferry, J.G. (2001). *Carbonic Anhydrase: New Insights for an Ancient Enzyme.* The Journal of Biological Chemistry, **276**: 48615-48618.

Ueda, T., Pichersky, E., Malik, V.S. and Cashmore, A.R. (1989). *Level of Expression of the Tomato* rbcS-3A *Gene Is Modulated by a Far Upstream Promoter Element in a Developmentally Regulated Manner.* The Plant Cell, **1**: 217-227.

Utsunomiya, E. and Muto, S. (1993). *Carbonic anhydrase in the plasma membrane of C_3 and C_4 plants.* Physiologia Plantarum, **88**: 413-419.

Venter, J.C. *et al.* (2001). *The Sequence of the Human Genome.* Science, **291**: 1304-1351.

Vidgren, J., Liljas, A. and Walker, N.P. (1990). *Refined structure of the acetazolamide complex of human carbonic anhydrase II at 1.9 Å.* International Journal of Biological Macromolecules, **12**: 342-344.

Villarejo, A., Shutova, T., Moskvin, O., Forssen, M., Klimov, V.V. and Samuelsson, G. (2002). *A photosystem II-associated carbonic anhydrase regulates the efficiency of photosynthetic oxygen evolution.* The EMBO Journal, **21**: 1930-1938.

Vullo, D., Voipio, J., Innocenti, A., Rivera, C., Ranki, H., Scozzafava, A., Kaila, K. and Supuran, C.T. (2005). *Carbonic anhydrase inhibitors. Inhibition of the human cytosolic isozyme VII with aromatic and heterocyclic sulfonamides.* Bioorganic and Medicinal Chemistry Letters, **15**: 971-976.

Walker, R.P., Chen, Z.H., Acheson, R.M. and Leegood, R.C. (2002). *Effects of phosphorylation on phosphoenolpyruvate carboxykinase from the C_4 plant Guinea grass.* Plant Physiology, **128**: 165-172.

Weiner, H., Burnell, J.N., Woodrow, I.E., Heldt, H.W. and Hatch, M.D. (1988). *Metabolite Diffusion into Bundle Sheath Cells from C_4 Plants.* Plant Physiology, **88**: 815-822.

Whitelaw, C.A., Barbazuk, W.B., Pertea, G., Chan, A.P., Cheung, F., Lee, Y., Zheng, L., van Heeringen, S., Karamycheva, S., Bennetzen, J.L., SanMiguel, P., Lakey, N., Bedell, J., Yuan, Y., Budiman, M.A., Resnick, A., Van Aken, S., Utterback, T., Riedmuller, S., Williams, M., Feldblyum, T., Schubert, K., Beachy, R., Fraser, C.M. and Quachenbush, J. (2003). *Enrichment of Gene-Coding Sequences in Maize by Genome Filtration.* Science, **302**: 2118-2120.

Wu, M.-X. and Wedding, R.T. (1994). *Modification of Maize Leaf Phosphoenolpyruvate Carboxylase with Fluorescein Isothiocyanate*. Plant Cell Physiology, **35**: 569-574.

Wyrich, R., Dressen, U., Brockmann, S., Streubel, M., Chang, C., Qiang, D., Paterson, A.H. and Westhoff, P. (1998). *The molecular basis of C_4 photosynthesis in sorghum: isolation, characterization and RFLP mapping of mesophyll- and bundle-sheath-specific cDNAs obtained by differential screening.* Plant Molecular Biology, **37**: 319-335.

Xu, Y., Feng, L., Jeffrey, P.D., Shi, Y. and Morel, F.M. (2008). *Structure and metal exchange in the cadmium carbonic anhydrases of marine diatoms.* Nature, **452**: 56-61.

Yanagisawa, S. (2000). *Dof1 and Dof2 transcription factors are associated with expression of multiple genes involved in carbon metabolism in maize.* The Plant Journal, **21**: 281-288.

Yanagisawa, S. (2004). *Dof Domain Proteins: Plant-Specific Transcription Factors Associated with Diverse Phenomena Unique to Plants.* Plant Cell Physiology, **45**: 386-391.

Yanagisawa, S., Akiyama, A., Kisaka, H., Uchimiya, H. and Miwa, T. (2004). *Metabolic engineering with Dof1 transcription factor in plants: Improved nitrogen assimilation and growth under low-nitrogen conditions.* Proceedings of the National Academy of Science, **101**: 7833-7838.

Yang, S.-Y., Tsuzuki, M. and Miyachi, S. (1985). *Carbonic Anhydrase of Chlamydomonas: Purification and Studies on its Induction Using Antiserum against Chlamydomonas Carbonic Anhydrase*. Plant Cell Physiology, **26**: 25-34.

Yu, S., Xia, D., Luo, Q., Cheng, Y., Takano, T. and Liu, S. (2007). *Purification and characterization of carbonic anhydrase of rice (*Oryza sativa *L.) expressed in* Escherichia coli. Protein Expression and Purification, **52,** 379-383.

Zhang, Y., Fan, W., Kinkema, M., Li, X. and Dong, X. (1999). *Interaction of NPR1 with basic leucine zipper protein transcription factors that bind sequences required for salicylic acid induction of the PR-1 gene.* Proceedings of the National Academy of Science, **96**: 6523-6528.

Zhang, J., Zhang, F. and Peterson, T. (2006). *Transposition of Reversed* Ac *Element Ends Generates Novel Chimeric Genes in Maize.* PLOS Genetics, **2**: 1535-1540.

Zhou, J., Tang, X., Frederick, R. and Martin, G. (1998). *Pathogen Recognition and Signal Transduction by the Pto Kinase.* Journal of Plant Research, **111**: 353-356.

Appendix - Chapter 3

3.1: Nucleotide sequences of the probes used for Southern blotting

Probe 1: Insert

```
GGCGGGCATAAGAGGGGAGGAGAGAGAGGCCAGAGAAGAGGAGGAGAAGAAGAAGAAGATGAGCAGCTGCCTCTGC
CTTCCGAAAAAAAAGGAGGGGCCAGCGAAGGAGAAGCCGTCCACAGATACCCCCACCTCGTCACTCCTTCAGAACC
AGAAGCCCTCCAACCTCCACCTCCTCCCTCCAAGGCTTCCTCCAAGGGC
```

Probe 2: Repeat

```
ATGGTGTTCGCCTGCTCCGACTCCCGCGTGTGCCCGTCGGTGACCCTGGGACTGCAGCCCGGCGAGGCATTCACCG
TCCGCAACATCGCTTCCATGGTCCCACCCTACGACAAGATCAAGTACGCCGGCACAGGGTCCGCCATCGAGTACGC
CGTGTGCGCGCTCAAGGTGCAGGTCATCGTGGTCATTGGCCACAGCTGCTGCGGTGGCATCAGGGCGCTCCTCTCC
CTCAAGGACGGCGCGCCCGACAACTTCCACTTCGTGGAGGACTGGGTCAGGATCGGCAGCCCTGCCAAGAACAAGG
TGAAGAAAGAGCACGCGTCCGTGCCGTTCGATGACCAGTGCTCCATCCTGGAGAAGGAGGCCGTGAACGTGTCGCT
CCAGAACCTCAAGAGCTACCCCTTCGTCAAGGAAGGGCTGGCCGGCGGGACGCTCAAGCTGGTTGGCGCCCACTAC
GACTTCGTCAAAGGGCAGTTCGTCACATGG
```

3.2: Nucleotide sequence of the probe used for screening the maize genomic DNA library

The polylinker sequence of the vector, which included the *EcoRI* recognition site, was removed from the 5′-end of this sequence so that only the CA2 cDNA sequence up to the *XhoI* recognition site is shown.

```
ATGTACACATTGCCCGTCCGTGCCACCACATCCAGCATCGTCGCCAGCCTCGCCACCCCCGCGCCGTCCTCCTCCT
CCGGCTCCGGCCGCCCCAGGCTCAGGCTCATCCGGAACGCCCCCGTCTTCGCCGCCCCCGCCACCGTCTGTAAACG
GGACGGCGGGCAGCTGAGGAGTCAAACGAGAGAGATCGAGAGAGAAAGAAAGGGAGGGCATCCACCAGCCGGCGGG
CATAAGAGGGGAGGAGAGAGAGGCCAGAGAAGAGGAGGAGAAGAAGAAGAAGATGAGCAGCTGCCTCTGCCTTCCG
AAAAAAAAGGAGGGGCCAGCGAAGGAGAAGCCGTCCACAGATACCCCCACCTCGTCACTCCTTCAGAACCAGAAGC
CCTCCCAACCTCCACCTCCTCCCTCCAAGGCTTCCTCCAAGGGCATGGACCCCACCGTCGAGCGCTTGAAGAGCGG
GTTCCAGAAGTTCAAGACCGAGGTCTATGACAAGAAGCCGGAGCTGTTCGAGCCTCTCAAGTCCGGCCAGAGCCCC
AGGTACATGGTGTTCGCCTGCTCCGACTCCCGCGTGTGCCCGTCGGTGACACTGGGACTGCAGCCCGGCGAGGCAT
TCACCGTCCGCAACATCGCTTCCATGGTCCCACCCTACGACAAGATCAAGTACGCCGGCACAGGGTCCGCCATCGA
GTACGCCGTGTGCGCCCTCAAGGTGGAGGTCCTCGTGGTCATTGGCCATAGCTGCTGCGGTGGCATCAGGGCGCTC
CTCTCCCTCCAGGATGGCGCACCTGACACCTTCCACTTCGTCGAGGACTGGGTTAAGATCGGCTTCATTGCCAAGA
TGAAGGTAAAGAAAGAGCACGCCTCGGTGCCGTTCGATGACCAGTGCTCCATTCTCGAG
```

3.3: Nucleotide sequence and alignments of the assembly obtained from the genomic DNA library screen with the CA2 gene.

This alignment was made between the CA2 gene sequence and the first assembly of 3.55 kb obtained from screening the maize genomic DNA library (Fig. 3.7).

A. ClustalW alignment of the end of Intron 1 of the CA2 gene with the genomic DNA library clone.

```
CA2 Gene             2651 CAGTTGAATAACCGACGACCATCAAATAAAAGGCCGCCACTACTAGTGGC 2700
Gen lib Clone 1         1 ------------------------------------------CTAGTGGC    8
                                                                    ********

CA2 Gene             2701 CATCGACGTCAGTTTAACCTTTCTATGTATGCATGTGTAACTTCCCATGA 2750
Gen lib Clone 1         9 CATCGACGTCAGTTTCACCTTTCTATGTATGCATGTGTAACTTCCCATGA   59
                          *************** **********************************

CA2 Gene             2751 TTTCCTTGGCTGCGTTATTTTGCTTTGTTTCACCGTCGGACGACGAAGTC 2800
Gen lib Clone 1        60 TTTCCTGCGTCGCGTTATTTTGCTTTGTTTCACCGTCGGTCGACGAAGTC  109
                          ******  *  **************************** **********

CA2 Gene             2801 TTTTAGATAG--CAATAAGGAACTATATCTAAGTCCTAGTTTGGGAACCT 2848
Gen lib Clone 1       110 TTTTAGTTAGAGCAATAAGGAACTATATCTAA------------------  141
                          ****** ***  ********************

CA2 Gene             2849 CGTTTTCCCACGAGATTTTCATTTTCCTAAGGTAAATTAGTTCCGGCTTT 2899
Gen lib Clone 1       142 --------CAGGAGATTT--------------------------------  151
                                  ** *******

CA2 Gene             2900 TTTGAAAATAAGAATCTTTTGAAAAAGATGGTAATTATCAAACTAGTCCT 2949
Gen lib Clone 1           --------------------------------------------------


CA2 Gene             2950 AACAGAGAGATTTTTGAGGGGGGAGAAAAAAAAGGAAGTTCTTCTGCA-- 2997
Gen lib Clone 1       152 ----------------GAGGGGGAGCTAGAAAAGGGAGTTCTTCTGCCAT  185
                                            *******  * ****** ***********

CA2 Gene             2998 ---TTCTTTTTTGGAGGAACAAAAAATTTGCCTCTGCATACTGAATCAGA 3044
Gen lib Clone 1       186 TCTTTTTTTTTTTGAGGAACAAAAAATTTGCCTCTGCATACTGAATCAGG  235
                             ** ****** ************************************

CA2 Gene             3045 GG-GGATGGGCTTTATTTCGTGTTGGCTGGTTGATTGATGATTGGATGAG 3093
Gen lib Clone 1       236 GGCGGGCGGGCTTTATTTCGTGTTGGCTGGTTGATTGACGATTGGATGAG  285
                          ** **  ******************************* ***********

CA2 Gene             3094 CTCCAGTAAGTTTGGAAGAGAACAGGGCACGGTCCCGACGGTTGGT---- 3139
Gen lib Clone 1       286 CTCCAGCAAGTTAGGAAGAGAACAGGGCACGGTCCCGACGGTTGGTTGGT  335
                          ****** ***** *********************************

CA2 Gene             3140 ACGGGTGAAGAAAGGGAGTGATTTAATTTATCGCCC-AACCACAACCACC 3188
Gen lib Clone 1       336 ACGGGTGAAGAAAGGGAGTGATTTAATTTATCGCCCCAACCACAACCATC  385
                          ************************************ *********** *

CA2 Gene             3189 CATCGATC-TATAGTTGCAGAAGAACTCGCTAATGCTGTC-----CACAA 3232
Gen lib Clone 1       386 CATCGATCCTATAGCTGCAGAAGAACTCGCTAATCTTGTCTTGTCCACAG  435
                          ******** ***** *******************  ****     ****

CA2 Gene             3233 AAGCCGCACTCACGCACTCATCCGCCACTGATTTTATTTCCCCCCCCCCC 3282
Gen lib Clone 1       436 AAGCCGCTCTCACGCACTCATCCGCCACTGATTTTATTTTTCCACCT---  482
                          ******* *******************************  ** **
```

```
CA2 Gene            3283 CCTGTGGCGCGCGGTTGCTGCGTGGTGGTACTACTACCTGTTTTTGCTCA 3332
Gen lib Clone 1      483 ------GTGCGCGCGTGCTGCGTGGTGGTACTACTACCTGTTTCT-CTCA  525
                               * *****  **************************** * ****

CA2 Gene            3333 CTGACACAGTTGCGGGT-TCATCATGTTGCT 3362
Gen lib Clone 1      526 CTGACACAGTTGCGCGCGTCATCATGTTGCT  556
                         ************** *  *************
```

B. Alignment of the genomic library clone with the CA2 gene and Exon 2 of the CA2 cDNA sequence (Burnell and Ludwig, 1997; gi:606810). Differences in the sequences are indicated in bold. The consensus translated sequence is also shown.

```
CA2 Gene    AGT AAA CGG GAC GGC GGG CAG CTG AGG AGT CAA ACG AGA GAG --- ATC GAG AGA GAA
Glib        AGT AAA CGG GAC GGC GGG CAG CTG AGG AGT CAA ACG AGA GAG -AG ATC GAG AGA GAA
gi606810    TGT AAA CGG GAC GGC GGG CAG CTG AGG AGT CAA ACG AGA GAG --- ATC GAG AGA GAA
             C   K   R   D   G   G   Q   L   R   S   Q   T   R   E       I   E   R   E

CA2 Gene   AGA AAG GGA GGG CAT CCA ---- CCA GCC GGC GGG CAT AAG AGG GGA GGA GAG AGA GGC
Glib       AGA AAG GGA GGG CAT CCA TCCA CCA GCC GCC GGC GAT AAA AGG GGA GGA GAG AGA GGC
gi606810   AGA AAG GGA GGG CAT CCA ---- CCA GCC GGC GGG CAT AAG AGG GGA GGA GAG AGA GGC
            R   K   G   G   H   P        P   A   G   G   H   K   R   G   G   E   R   G

CA2 Gene    CAG AGA AGA GGA GGA GAA GAA GAA GAA GAT GAG CAG CTG CCT CTG CCT TCC GAA AAA
Glib        CAG AGA AGA GGA GGA GAA GAA GAA GAA GAT GAG CAG CTG CCT CTG CCT TCC GAA AAA
gi606810    CAG AGA AGA GGA GGA GAA GAA GAA GAA GAT GAG CAG CTG CCT CTG CCT TCC GAA AAA
             G   R   R   G   G   E   E   E   E   D   E   Q   L   P   L   P   S   E   K

CA2 Gene    AAA GGA GGG GCC AGC GAA GGA GAA GCC GTC CAC AGA TAC CCC CAC CTC GTC ACT CCT
Glib        AAA GGA GGG GCC AGC GAA GGA GAA GCC GTC CAC AGA TAC CCC CAC CTC GTC ACT CCT
gi606810    AAA GGA GGG GCC AGC GAA GGA GAA GCC GTC CAC AGA TAC CCC CAC CTC GTC ACT CCT
             K   G   G   A   S   E   G   E   A   V   H   R   Y   P   H   L   V   T   P

CA2 Gene   TCA GAA CCA GAA GCC CTC C-AA CCT CCA CCT CCT CCC TCC AAG GCT TCC TCC AAG GGC
Glib       TCA GAA CCA GAA GCC CTC CCAA CCT CCA CCT CCT CCC TCC AAG GCT TCC TCC AAG GTC
gi606810   TCA GAA CCA GAA GCC CTC C-AA CCT CCA CCT CCT CCC TCC AAG GCT TCC TCC AAG GGC
            S   E   P   E   A   L    Q   P   P   P   P   P   S   K   A   S   S   K   G
```

C. ClustalW alignment of Intron 2 of the CA2 gene with the genomic DNA library clone.

```
CA2 Gene               1 CGTCCCCTCCTCCTCCTCCTCATCTTCCTCTCTCACCTTCAGCACCATCC   50
Gen lib Clone 1        1 CGTCCCCTCCTCCTCCTCCTCCACCTCCTCCTCCTCTCTCACCTTCAG--   48
                         *********************  * *****   * *  *** *  **

CA2 Gene              51 TCCACACAGCAGCACGCGCGCAGCAATCTCACCGTTTTCTTTTCCTCCAT  100
Gen lib Clone 1       49 --CACGTACCATCCTCCACCCCG-----TTTCCCTTTTCTTTTCCTCCAT   91
                         * ***  * ** *   * * * *     *  ** ****************

CA2 Gene             101 TGCCATCAGTAGCTAGCCACACTGCATGCATTCAGCTTCCGCTTTCTCCC  150
Gen lib Clone 1       92 TGCCATCAGTAGCTAGCCACACTGCATGCATTCAGCTTCCGCTTTCTCCC  141
                         **************************************************

CA2 Gene             151 TGTGTAGCGAGCGCTGGTGCCGGCCGGTGCAGAGAAGATCCCTGCTCCCC  200
Gen lib Clone 1      142 TGAGTAGCGAGCGCTG-TGCCGGCCGGTGCAGAGAAGATCCCTGTTCCCA  190
                         ** ************* *************************** ****

CA2 Gene             201 CCCCCCCCCCCCCCCCCTAATTAAGATCACCTTTGTGCATTTTTTTCC-T  249
Gen lib Clone 1      191 CCCCCTGCCCCT--------TTTAGATCACCTTTGTGCATTTTTTTCCCT  232
                         *****  ****         ** ************************* *
```

```
CA2 Gene               250 TGTGTTGTGGTCCGTCGGCAAGTAGGCCAAAATTGCATCATGC------C  293
Gen lib Clone 1        233 TGTGTTGTGGTCCGTCGGCAAGGAGGCCAAAATTGCATCATGCATGGGCC  282
                           ********************** ********************      *

CA2 Gene               294 ATGGCCCCTCCTCTT-CTACTACCTCGTCATGCAGC--CAGCAACGACAT  340
Gen lib Clone 1        283 ATGGCCCCTCCTCTTTCTACTACCTCGTCATGCTGCATCACGATCATGAT  332
                           *************** ***************** **  **  * *   **

CA2 Gene               341 GAATGACC------------------------------------------  348
Gen lib Clone 1        333 GAATGAGCATGCATCCAATCCAAGTTGGTCTCTCTGGGGATGATGGCGAG  382
                           ****** *

CA2 Gene               349 -----------CGAACGAAGTATCTGGC-GTTGACATTGCAG  378
Gen lib Clone 1        383 AAACTGATGACCCGACGATGTATCTGGCTGTTGACATTGCAG  424
                                      *  **** ********* *************
```

D. Alignment of the genomic DNA library clone with the CA2 gene and Exon 3 of the CA2 cDNA sequence (Burnell and Ludwig, 1997; gi:606810). Differences in the sequences are indicated in bold. The consensus translated sequence is also shown.

```
CA2 Gene    GGC ATG GAC CCC ACC GGC GAG CGC TTG AAG AGC GGG TTC CAG AAG TTC AAG ACC GAG
Glib        GGC ATG GAC CCC ACC GTC GAG CGC TTG AAG AGC GGG TTC CAG AAG TTC AAG ACC GAG
gi:606810   GGC ATG GAC CCC ACC GTC GAG CGC TTG AAG AGC GGG TTC CAG AAG TTC AAG ACC GAG
             G   M   D   P   T   V   E   R   L   K   S   G   F   Q   K   F   K   T   E

CA2 Gene      GTC TAT GA
Glib          GTC TAT GA
gi:606810     GTC TAT GA
               V   Y  (D)
```

E. ClustalW alignment of Intron 3 of the CA2 gene with the genomic DNA library clone.

```
CA2 Gene                 1 GTAAGTCACCTGAGCTGTTTGTTCTCTGCAGCAACCCGCGTTTGGTTTCT   50
Gen lib Clone 1          1 GTAAGTCACCTGAGCT----GTTCTCTGCAGCA-CCCGCGTTTGGTTTCT   45
                           ****************    ************* ****************

CA2 Gene                51 ATTTCCTTTTTTTTTGTTTGTGAATTCAGTGAGCTCCGACTCCGACTGAT  100
Gen lib Clone 1         46 GTTTCCTTTTTTTT-----GTGAATTCAGTGAGCTCCGACTCCGACTGAT   90
                            *************     *******************************

CA2 Gene               101 TCATGTGCTCCGCTGATCTTTGTTCGCAG  129
Gen lib Clone 1         91 TCCCGTGCTCCGCTGATCTTTGTTCGCAG  119
                           **  *************************
```

F. Alignment of the CA2 gene, the genomic DNA library clone and Exon 4 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810). There are no differences in the three sequences. The consensus translated sequence is also shown.

```
CA2 Gene      --C AAG AAG CCG GAG CTG TTC GAG CCT CTC AAG TCC GGC CAG AGC CCC AG-
Glib          --C AAG AAG CCG GAG CTG TTC GAG CCT CTC AAG TCC GGC CAG AGC CCC AG-
gi:606810     --C AAG AAG CCG GAG CTG TTC GAG CCT CTC AAG TCC GGC CAG AGC CCC AG-
                   K   K   P   E   L   F   E   P   L   K   S   G   Q   S   P   R
```

G. ClustalW alignment of Intron 4 of the CA2 gene with the genomic DNA library clone.

```
CA2 Gene                     1 GGTATGCGCTGCTAATGTTTTTTTATATATATTTTGTTGTGTGTCTATAG   50
Gen lib Clone 1              1 GGTATGCGCTGCTAATGTTTTTTTATATATATTTTGTTGTGTGTCTCTAG   50
                               ********************************************** ***

CA2 Gene                    51 CGACTCCGGGCAACTGGGGCAAAAGATTGAAGTAGTACTAGTTGCTCGTT  100
Gen lib Clone 1             51 CGACTCCGGCCAACTGGGCCAAAAGATTGA-GTAGTGCTAGTTGCTCGTT   99
                               ********* ******** *********** ***** *************

CA2 Gene                   101 CCTATTACTAGCTCTGTAGCTCATCACCATTGCTGC---------TGCAA  140
Gen lib Clone 1            100 CCTATTACTAGCTCTGTAGCTCATCACCATTGCTGCATTCCCGTCTGTAA  149
                               ************************************         ** **

CA2 Gene                   141 CACCCTGCCGCA--------CCTGCAC------TATTCAGCATC---CAC  163
Gen lib Clone 1            150 CTGTAAGACGCAAGAGTAGGCCTACGCAAGAGGCAAGCAACACCGTGCAC  199
                               *     * ****        *** * *       *  ** ** *   ***

CA2 Gene                   164 CCTGTCTCCCCTGGACCAAAGCTGCAAG-----GGGAACCATGCAGATAA  208
Gen lib Clone 1            200 TGTGTCTCCCCTGGGCCAAAGCTGCAAGCTAAGGTGAACCATGCAGATAA  249
                                 ************ *************     * ***************

CA2 Gene                   209 TACTAGGTGTGTATTATCAGCATTCCATGGCTAATGTGTGGTCCAGGCGT  258
Gen lib Clone 1            250 TACTAGGTGTGTATTATCAGCATTCCATGGCTAATGTGTGGTCCAG----  295
                               **********************************************

CA2 Gene                   259 CCAGCACTGTGGGTCGCCCCAC-TCACGGGATCCTGTCGTCATCGTGAGT  307
Gen lib Clone 1            296 ----CACTGTCCCTCGCCCCACCTCACGGGATCCTTTCGTCATCGTGAGT  341
                                   ******   ********* ************ **************

CA2 Gene                   308 AGTTGGCTTGGACGTGTCCCCTTCCCCTCTCGCACCCCTTGCAAAAAAGT  357
Gen lib Clone 1            342 AGTTGGCTTGGACGTGTCCCCTCTCACACCC-CGGCCCTTGCAAAAAAGT  390
                               **********************  * * * * *  ***************

CA2 Gene                   358 TAGGTGCATAAATGTTGGGCCTGTTGCCGGTCCTCGAGGAAATATGCTAC  407
Gen lib Clone 1            391 TAGGTGCATAAATGTTGGGCCGGTTGCAGGTCCTCGAGGAAATATGCTAC  440
                               ********************* ***** **********************

CA2 Gene                   408 ACTACAGATGTCCCAATTTTTGTGGAAGATATGGCAGCAGCATCACGCCT  457
Gen lib Clone 1            441 ACTACAGATGTCCCAATTTTGGTGGAAGATG-----GCAGCATCACGCCT  485
                               ******************** *********      **************

CA2 Gene                   458 CCTGATGATGCCCGGAACGGAAATGTTCTTGCTATTGGCCGCCAGCAGGG  497
Gen lib Clone 1            486 CCTGA---TGCCCGGAACGGAAATGTTCTTGCTATTGGCCGCGCGCCAGG  532
                               *****   **********************************  **  **

CA2 Gene                   498 GAATATAATGGGATAAAGATAGACCAGCGTGCTAGAGAGCCACACGGAAA  547
Gen lib Clone 1            533 GAATATAATGGGATGAAGATAGGCCAGCGTGCTAGAGAGCCACACGGAAA  582
                               ************** ******* ***************************

CA2 Gene                   548 CCAGAGCGCGCGTAGAGCATCCTCGTCGCAACT-------AATACTAGTA  590
Gen lib Clone 1            583 CCAGAGCGCGCGTAGAGCATCCTCGTCGCAACTCGCAACTACTACTAGTA  632
                               *********************************       * ********

CA2 Gene                   591 CTTACAGAGCCAGAGGAGGAGGGTCAAATCGAAACTCAATCAAAAGCTTG  640
Gen lib Clone 1            633 CTTACAGAGACAGAGGAGG---GTCAAATCGAAACTGAATCAAAAGCTTG  679
                               ********* *********   ************** *************

CA2 Gene                   641 TCGCCTTTTTGGGGCGCCAGAAATCTTCCACTGATGAGATGACCAGGGCC  690
Gen lib Clone 1            680 CCGCCTTTTTGGGGCGCCAGAAATCCTCCACTCCACTGATGACCAGGGCC  729
                                ************************ ******     *************

CA2 Gene                   691 GATGATCTGCTTACCTGCTTATCGATAAGAGCCATGGGAAACCGATCGAA  740
Gen lib Clone 1            730 GATGATCTGCTTACCTGCTTATCGATAAGAGCCATGTGAC-CAGATCGAA  778
                               ************************************ **  * *******
```

```
CA2 Gene                  741 CTTGGTTTTGCGTACGTGCTCCTCCCTCTTTTCACCGACTG-----ACGG  785
Gen lib Clone 1           779 CTTGGTTTTGCGTACGTGCTCCTCCCTCTTTTCGCCGACCTGACGCACGG  828
                              ********************************* *****       ****

CA2 Gene                  786 GTGACTGATTTCCCCTC---------------CGCTGCAG   810
Gen lib Clone 1           829 GTGACTGATTTCCCTCCGGCCGCTGCTGGCATTGCTGCAG   868
                              **************  *                *******
```

H. Alignment of the CA2 gene, the genomic DNA library clone and Exon 5 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810). There are no differences in the three sequences. The consensus translated sequence is also shown.

```
CA2 Gene   --G TAC ATG GTG TTC GCC TGC TCC GAC TCC CGC GTG TGC CCG TCG GTG ACC CTG GGA
Glib       --G TAC ATG GTG TTC GCC TGC TCC GAC TCC CGC GTG TGC CCG TCG GTG ACC CTG GGA
gi:606810  --G TAC ATG GTG TTC GCC TGC TCC GAC TCC CGC GTG TGC CCG TCG GTG ACC CTG GGA
                Y   M   V   F   A   C   S   D   S   R   V   C   P   S   V   T   L   G

CA2 Gene   CTG CAG CCC GGC GAG GCA TTC ACC GTC CGC AAC ATC GCT TCC ATG GTC CCA CCC TAC
Glib       CTG CAG CCC GGC GAG GCA TTC ACC GTC CGC AAC ATC GCT TCC ATG GTC CCA CCC TAC
gi:606810  CTG CAG CCC GGC GAG GCA TTC ACC GTC CGC AAC ATC GCT TCC ATG GTC CCA CCC TAC
            L   Q   P   G   E   A   F   T   V   R   N   I   A   S   M   V   P   P   Y

CA2 Gene     GAC AAG
Glib         GAC AAG
gi:606810    GAC AAG
              D   K
```

I. ClustalW alignment of Intron 5 of the CA2 gene with the genomic DNA library clone.

```
CA2 Gene                    1 GTACGT-ACGTACGAGCAAACACCGATCGACGCATGCAACGGTGGTAT--  47
Gen lib Clone 1             1 GTACGTTACGTACGAGCAAACACCGATCGACGCATGCAATGGTGTTTTTT  50
                              ****** ******************************** **** * *

CA2 Gene                   48 ----------------CAGCCACACTAATATT------------------  63
Gen lib Clone 1            51 TTTTCTTTTTCTGGCACGGGTATAAGGGTATTGGCAGAAGATGCATGGAT 100
                                              * *  * *    *  *

CA2 Gene                   64 -ACTCACACGGTCGTC----TTCCGTTTT-GGCCAAACTGCAG   96
Gen lib Clone 1           101 GACTCACACGGCGGTCGGTCTTCCGTTTTTGGCCAAACTGCAA  143
                               **********  ***    ********* ************
```

J. Alignment of the CA2 gene, the genomic DNA library clone and Exon 6 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810). Differences are indicated with bolding. The consensus translated sequence is also shown.

```
CA2 Gene   ATC AAG TAC GCC GGC ACC GGG TCC GCC ATC GAG TAC GCC GTG TGC GCG CTC AAG GTG
Glib       ATC AAG TAC GCC GGC ACA GGG TCC GCC ATC GAG TAC GCC GTG TGC GCG CTC AAG GTG
gi:606810  ATC AAG TAC GCC GGC ACA GGG TCC GCC ATC GAG TAC GCC GTG TGC GCG CTC AAG GTG
            I   K   Y   A   G   T   G   S   A   I   E   Y   A   V   C   A   L   K   V
```

```
CA2 Gene   CAG GTC ATC GTG GTC ATT GGC CAC AGC TGC TGC GGT GGC ATC AGG GCG CTC CTC TCC
Glib       CAG GTC ATC GTG GTC ATT GGC CAC AGC TGC TGC GGT GGC ATC AGG GCG CTC CTC TCC
gi:606810  CAG GTC ATC GTG GTC ATT GGC CAC AGC TGC TGC GGT GGC ATC AGG GCG CTC CTC TCC
            Q   V   I   V   V   I   G   H   S   C   C   G   G   I   R   A   L   L   S

AZM4_68973   CTC AAG GAC GGC GCG CCC GAC AAC TT
CA2 Gene     CTC AAG GAC GGC GCG CCA GAC AAC TT
gi:606810    CTC AAG GAC GGC GCG CCC GAC AAC TT
              L   K   D   G   A   P   D   N  (F)
```

K. ClustalW alignment of Intron 6 of the CA2 gene and the genomic DNA library clone.

```
CA2 Gene           1 GTAAGCAGTAGTCATCGTAAAATGCGTATAAAAAATATATATAGCAGTTT   50
Gen lib Clone 1    1 GTAAGCAGTAGTCATCGTAAAATGCGTATATAAATTATAT--AGCAGTTT   48
                     ****************************** *** *****  ********

CA2 Gene          51 TATTTAGAGAGAGAGAAAAAAATTAGAACCCCGTGTAGTGTAATGCTCAG  100
Gen lib Clone 1   49 TATTTAGCGAGAGGGAGGGAGAGAAAAA--------AATAGAACCC----   86
                     ******* ***** **   * *  * **        * *  **  *

CA2 Gene         101 CGTGTTGTCTGTCGTTGGTTTAAATCTGGCCATGTATATCCAG  143
Gen lib Clone 1   87 CGTGTTGTCT---GTTGGTTTGA-TTTGGCGATGTATATCCAG  125
                     **********   ******** * * *****************
```

L. Alignment of the CA2 gene, the genomic DNA library clone and Exon 7 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810). Differences are indicated with bolding. The consensus translated sequence is also shown.

```
CA2 Gene   --C CCA TTC GTG GAG GAC TGG GTC AGG ATC GGC AGC CCT GCC AAG AAC AAG GTG AAG
Glib       --C CAC TTC GTG GAG GAC TGG GTC AGG ATC GGC AGC CCT GCC AAG AAC AAG GTG AAG
gi:606810  --C CAC TTC GTG GAG GAC TGG GTC AGG ATC GGC AGC CCT GCC AAG AAC AAG GTG AAG
                H   F   V   E   D   W   V   R   I   G   S   P   A   K   N   K   V   K

CA2 Gene     AAA GAG CAC GCA TCG GTG CCG TTC GAT GAC CAG TGC TCC ATC CTG GAG AAG
Glib         AAA GAG CAC GCG TCG GTG CCG TTC GAT GAC CAG TGC TCC ATC CTG GAG AAG
gi:606810/4  AAA GAG CAC GCG TCC GTG CCG TTC GAT GAC CAG TGC TCC ATC CTG GAG AAG
              K   E   H   A   S   V   P   F   D   D   Q   C   S   I   L   E   K
```

M. ClustalW alignment of Intron 7 of the CA2 gene and the genomic DNA library clone.

```
CA2 Gene           1 GTACGTAACGTAA-ACGCACGCACACACACCGACCGTATGAATAATGGAT  49
Gen lib Clone 1    1 GTACGTAACGTAACGTAAACGCACGCACGCACACCGCATGAATAATGGAT  50
                     *************     ****** *** *  **** *************

CA2 Gene          50 TATATATTATTGGT---TTCGCTCATCAACGAACAAATTCAAGGATCATC  96
Gen lib Clone 1   51 TATATATTATTGGTAATTTCGCTCATCAACAAACAAATTAAAGGATCATC 100
                     **************   ************* ******** **********

CA2 Gene          97 ATCGACCTTTAATTGTGTGTGTGTTTCTGCAG  130
Gen lib Clone 1  101 ---GACCTTTAATTGC---TGGATGTTTCTGCAG  128
                        ************    **  ***********
```

N. Alignment of the CA2 gene, the genomic DNA library clone and Exon 8 of the CA cDNA sequence (corresponding to the last exon of Repeat A; Burnell and Ludwig, 1997; gi:606810 and gi:616814). Differences are indicated with bolding. The consensus translated sequence is also shown.

```
CA2 Gene   GAG GCC GTC AAC GTC TCG CTC CAG AAC CTC AAG AGC TAC CCC TTC GTC AAG GAA GGG
Glib       GAG GCC GTG AAC GTG TCG CTC CAG AAC CTC AAG AGC TAC CCC TTC GTC AAG GAA GGG
606810/4   GAG GCC GTG AAC GTG TCG CTC CAG AAC CTC AAG AGC TAC CCC TTC GTC AAG GAA GGG
            E   A   V   N   V   S   L   Q   N   L   K   S   Y   P   F   V   K   E   G

CA2 Gene   CTG GCC GGC GGG ACG CTC AAG CTG GTT GGC GCC CAC TAC GAC TTC GTC AAA GGG CAG
Glib       CTG GCC GGC GGG ACG CTC AAG CTG GTT GGC GCC CAC TAC GAC TTC GTC AAA GGG CAG
gi:606810  CTG GCC GGC GGG ACG CTC AAG CTG GTT GGC GCC CAC TCA GAC TTC GTC AAA GGG CAG
gi:606814  CTG GCC GGC GGG ACG CTC AAG CTG GTT GGC GCC CAC TAC AGC TTC GTC AAA GGG CAG
            L   A   G   G   T   L   K   L   V   G   A   H   Y   D   F   V   K   G   Q

CA2 Gene     TTC GTC ACA TGG GAG CCT
Glib         TTC GTC ACA TGG GAG CCT
gi:606810/4  TTC GTC ACA TGG GAG CCT
              F   V   T   W   E   P
```

3.4: Sequence alignment of AZM4_68974 and the CA2 gene.

A. ClustalW alignment of the AZM4_68974 assembly with the CA2 gene in the region (2.5 kb) upstream of the start codon (ATG).

```
AZM4_68974          1 CTACTGCTACCTTAAAATATGCATGCATGCACGACAAAGGGTAAACGCGC   50

AZM4_68974         51 ATGACGCCACCAGTTTAATTTGTAGCTTTTGCATTTTCAATTCCCTTACC  100

AZM4_68974        101 TTTGTGTTGTCGCCGTCACGTACGCGCTGTCTAGCTAGCGATCTCATGTG  150

AZM4_68974        151 CTTATGATGTCGCCAAATGAGCACATCCTGGCAGCATGCCAAGGCTCAAG  200

AZM4_68974        201 TGTGTCAACCTTGCACGTACATGTACGTCGCCGGTTYCGTTCCGTCCACC  250

AZM4_68974        251 GGGCAGCACAATGAGCTTCAGCTAGCGCTGGGCCGTGCAAACCCTAACCG  300

AZM4_68974        301 CCGCCCACCGCCATGGATCGGGCGGCATGCATAGAGTTCGTGGCGAGCGG  350

AZM4_68974        351 CGGAGCGCTAGGTAGAGCTTAGACCTCATCAGATTTGTTCTACTCTACTG  400

AZM4_68974        401 AATCCTGGTTGCATYTCATGGATGCATGTTTYTTCTTCTGCTTATTATTG  450

AZM4_68974        451 TCTGTGCTTCAAGCTCTGTCAATGCTATATATATATATGCTCCGCCACTG  500

AZM4_68974        501 CTCGTCGCTAATTCGATCGGCAGCGCGGGCATCGGAGCAGCTAGCCACGC  550

AZM4_68974        551 ATGCATCAGCCAAGCTTGGATGGAATGGAGCAGCTGGGACGGTTCCAGAT  600

AZM4_68974        601 GATTAAATATTATACTCCGTACAAATTATACAGCATCCCAAATGGATTCT  650

AZM4_68974        651 TTGTAACCGAACAAGTCAGGATGAATACGGCGGGCATCGGAGAAAGGTGA  700

AZM4_68974        701 GTTCATCCACTTCTCTCTCACAACTCTGTCAAGAGTGGTATATATATGTT  750

AZM4_68974        751 TAAACTATATTGCATATTCCAGTGAGACATGTCAGCTCTGGGGCGCTGCA  800

AZM4_68974        801 GCCTGGATCCAGAAGCTTCACAATTCTATACCACCCCTCACTTGTTCTGT  850

AZM4_68974        851 TGTTTGAAAACCAGGTTTTCGAACGACATTTCAAGCTCAATTAGCCCAGT  900

AZM4_68974        901 ACCCAGTTACCCCACTGGGGGACACTTTGTACATGAGCAGCCAGCCAATC  950

AZM4_68974        951 TGCGCATAGGCTAACGCCTAACGGCCTGGCCCCGCCCATGTCAGCAGGCC 1000

AZM4_68974       1001 TCCGAGGCTTTTGGTTGCCCAACCAGCCCATGGGCTGAATTCATAACAGT 1050
CA2 Gene            1                                     GAATTCATAACAGT   14
                                                          **************

AZM4_68974       1051 GTTGGCACACAGTTTCCTCTTCACTCGGAAGCTTATTATTATCGATCCTG 1100
CA2 Gene           15 GTTGGCACACAGTTTCCTCTTCACTCGGAAGCTTATTATTATCGATCCTG   64
                      **************************************************

AZM4_68974       1101 AACCAGAGACTAGCAGAGCTAGCATTTCGACGACGCGTCTCAACTCTCAA 1150
CA2 Gene           65 AACCAGAGACTAGCAGAGCTAGCATTTCGACGACGCGTCTCAACTCTCAA  114
                      **************************************************

AZM4_68974       1151 CCTCCAAGTCCACCTCGTGTACGTGCTGCCTTGCCAGTTGCCACTGGGCA 1200
CA2 Gene          115 CCTCCAAGTCCACCTCGTGTACGTGCTGCCTTGCCAGTTGCCACTGGGCA  164
                      **************************************************

AZM4_68974       1201 CTGCTGGCCCAGTGACCAACCATGCGTTAGATCTGACAGCACCACCGAAC 1250
CA2 Gene          165 CTGCTGGCCCAGTGACCAACCATGCGTTAGATCTGACAGGACCACCGAAC  214
                      *************************************** **********


AZM4_68974       1251 CATCCTCCCCGGTGATCAACAAACGACGGCAGCCACATCTTGCACCCAAC 1300
CA2 Gene          215 CATCCTCCCCGGTGATCAACAAACGACGGGAGCCACATCTTGCACCCAAC  264
                      ***************************** ********************

AZM4_68974       1301 GTGATGATGAATGATGCCTAGAACTTTTGACAACAAAACGCAGCACAGGT 1350
CA2 Gene          265 GTGATGATGAATGATGCCTAGAACTTTTGACAACAAAACGCAGCACAGGT  314
                      **************************************************
```

```
AZM4_68974      1351 AGCAGGTTTAATTCAACAAGACTTTCTACTATATAGAGCCACACCATAGA 1400
CA2 Gene         315 AGCAGGTTTAATTCAACAAGACTTTCTACTATATAGAGCCACACCATAGA  364
                     **************************************************

AZM4_68974      1401 GATAACTAATCTGTGCGCAAAGCCAAAGTGCTGAC----GGCAACTGTGG 1446
CA2 Gene         365 GATAACTAATCTGTGCGCAAAGCCAAAGTGCTGACTGACGGCAACTGTGG  414
                     ***********************************    ***********

AZM4_68974      1447 TGCAGCCTTTTCATCTCCGTTTTTAAGTTTTTTGCCCCTCCTTTTGTTTT 1496
CA2 Gene         415 TGCAGCCTTTTCATCTCCGTTTTTAAGTTTTTTGCCCCTCCTTTTGTTTT  464
                     **************************************************

AZM4_68974      1497 CTGTTTTTCTGGGAACTCTTTAAACCGCCGTGGCGCCGTGTAAACTTTGC 1546
CA2 Gene         465 CTGTTTTTCTGGGAACTCTTTAAACCGCCGTGGCGCCGTGTAAACTTTGC  514
                     **************************************************

AZM4_68974      1547 TGTAGCCTTTTCGCGTGCAATGGCAGAGCGCCCTGTTCTTTTCCTGCTAA 1596
CA2 Gene         515 TGTAGCCTTTTCGCGTGCAATGGCAGAGCGCCCTGTTCTTTTCCTGCTAA  564
                     **************************************************

AZM4_68974      1597 AGAA--AAAAAAAAAGGAGCACCTGATCGCTGGCAGGCCCACGGCCCACC 1644
CA2 Gene         565 AGGAGAAAAAAAAAAGGAGCACCTGATCGCTGGCAGGCCCACGGCCCACC  614
                     ** *  ********************************************

AZM4_68974      1645 CAACTGTGTCTGTAACGCTCGGCGTCCCTGCATTGCATGCCAAGTGCCAA 1694
CA2 Gene         615 CAACTGTGTCTGTAACGCTCGGCGTCCCTGCATTGCATGCCAAGTGCCAA  664
                     **************************************************

AZM4_68974      1695 CCACCAGTCCATAGCAGGGTCAGGGAGACCGCAGATGAGGCCGGGGCAAC 1744
CA2 Gene         665 CCACCAGTCCATAGCAGGGTCAGGGAGACCGCAGATGAGGCCGGGCCAAC  714
                     ********************************************* ****

AZM4_68974      1745 GGTGATGCCGCAAAGAGGATTCAGAATCCTTTTTCTTTTCTTTTCTTTTA 1794
CA2 Gene         715 GGTGATCCCCCAAAGAGGATTCAGAATCCTTTTTCTTTTCTTTTCTTTTA  764
                     ****** ** ****************************************

AZM4_68974      1795 CCACCGGGCTGGCATCACAGATTACACGCGCAGTAGAGTAAGCACGTCTC 1844
CA2 Gene         765 CCACCGGGCTGGCATCACAGATTACACGCGCAGTAGAGTAAGCACGTCTC  814
                     **************************************************

AZM4_68974      1845 TCTCGTAGCCAAGAACAACAGTCTA-CACAGCTCGCTTTCTCCGCCCTTG 1893
CA2 Gene         815 TCTCGTAGCCAAGAAAAACAGTCTAGCACAGCTCGGATTCTCC-CCCTTG  863
                     *************** ********* *********  ****** ******

AZM4_68974      1894 TCTGGGCGTTACGGCAGGCAAGCCCCCTCGTTTTCTTCTGCTCGCGTTCT 1943
CA2 Gene         864 TCTGGGCGTTACGGCAGGCAAGCCCCCTTGTTTTCTTCTGCTGGCGTTCT  913
                     **************************** ************* *******

AZM4_68974      1944 CCTTCCATGTCCACATCTCCTGTGCCACCGCACGCAAGGTGCCAACGCTC 1993
CA2 Gene         914 CCTTCCATGTCCACATCTCCTGTGCCACCGCACGCAAGGTGCCAACGCTG  963
                     ************************************************* 

AZM4_68974      1994 CCTCGCCGCAGTAGCATCGCGTCCACACAAACTGCACCTCCACTAGATAC 2043
CA2 Gene         964 CCTCGCCGCAGTAGCATCGCGTCCACACAAACTGCACCTCCACTAGATAC 1013
                     **************************************************

AZM4_68974      2044 GGCGGTGATCCGGCGAGAGAGCGCGACACGCACAGGCCAGCTAGCGTTTC 2093
CA2 Gene        1014 GGCGGTGATCCGGCGAGAGAGCCCGATACGCACAGAACAGCTAGCGTTTC 1063
                     ********************** *** ********  *************

AZM4_68974      2094 TCC--GACGCCGCGCGTTTCATCATTTCCCGCTTCCCCTGCCCCCGGCCG 2141
CA2 Gene        1064 CTCACGACGCCGCGCGTTTCATCATTTCCCGCTTCCACTGCCTCCGGCNG 1113
                       *  ******************************* ***** ***** *

AZM4_68974      2142 CG--CGCGCGCGCCCGTGTGGTCCAGACCAGGACGCGCGCGGATGTGCAT 2189
CA2 Gene        1114 NGNGCGCGCGCGCCCGTGTGGTCCAGAACCAGGAGCCCG-GGATGTGCAT 1162
                      *  *********************** *  *  ** ** **********

AZM4_68974      2190 CCGGCGCGCGCCCGTCGGCCACACGGTGCCGCCGCGCGTTATCCCGAGCC 2239
CA2 Gene        1163 CCGGCGCGC-CCCGTCGGCAACACGGTACCGCCGCGCGTTATCCCGAACC 1211
                     ********* ********* ******* ******************* **

AZM4_68974      2240 CTGTCCTGTCCTGTCCTGTTCCATCTCGCGCGCGAGGGGGGGAGGGGAGG 2289
CA2 Gene        1212 CTGTCCTGTCCTGTCCTGTTCCATCTCGCGCGCGAGGGGGGGAGGGGAGG 1261
                     **************************************************

AZM4_68974      2290 GCAGCGAGTGGCGCGCTGGCGGATGAGGCGCCGAGTGGCCCGCATCCACC 2339
```

```
CA2 Gene        1262 TCAGCGAGTGGCACGCTGGCGGATGAGGCGCCGAGGTGCCCTCATCTACC 1311
                      *********** *********************   **** **** ***

AZM4_68974      2340 GGCGCAGGCGAGCCGCACGACGCCGCCGCGCTCGCGGACCGCCGCCGCCA 2389
CA2 Gene        1312 GTCGCAGGCGATCCGCACGAAGCCGCCGCGCTCGCG-ACCGCCGCCGCCA 1360
                     * ********* ******** *************** *************

AZM4_68974      2390 CACATGCGCACCCCCGGCCCGCGGGGCTGTAACGGCCTTGTCGCCACGCG 2439
CA2 Gene        1361 CACATGCGCACCCCCGGCC-GCGGGGCTGTAACGGCCTTGTCGCCACGCG 1409
                     ******************* ******************************

AZM4_68974      2440 TGCGCCCCGTGTGTATAAGGAGGCAGCGCGTACAGGGGGCACGATAAGC- 2488
CA2 Gene        1410 TGCGCCCCGTGTGTATAAGGAGGCAGCGCGTACAGGGGGCACGATAAGCC 1459
                     *************************************************

AZM4_68974      2489 -------------------------------------------------- 2489
CA2 Gene        1460 TTGTCACNANGCGTGCNCCCCGTTTGAATAAGGAGGCAAGNGCGTACAGG 1509

                                                  Start
AZM4_68974      2489 -------------GGCACTCGCACGATCAATGTACACATTGCCCGTCCGC 2525
CA2 Gene        1510 GGGCACGATAAGCGGGACTCGCACGATCAATGTACACATTGCCCGTCCGT 1559
                                  ** *********************************
```

B. Alignment of AZM4_68974 and the CA2 gene with Exon 1 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810 and gi:616814). Differences in the sequences are indicated in bold. The consensus translated sequence is also shown.

```
4_68974   ATG TAC ACA TTG CCC GTC CGC GCC ACC ACA TCC AGC ATC GTC GCC AGC CTC GCC ACC
CA2 Gene  ATG TAC ACA TTG CCC GTC CGT GCC ACC ACA TCC AGC ATC GTC GCC AGC CTC GCC ACC
gi606814  ATG TAC ACA TTG CCC GTC CGT GCC ACC ACA TCC AGC ATC GT- GCC AGC CTC GCC ACC
gi606810  ATG TAC ACA TTG CCC GTC CGT GCC ACC ACA TCC AGC ATC GTC GCC AGC CTC GCC ACC
           M   Y   T   L   P   V   R   A   T   T   S   S   I   V   A   S   L   A   T

4_68974   CCC GCG CCG TCC TCC TCC TCC GGC TCC GGC --- --- CGC CCC AGG CCC AGG CTC ATC
CA2 Gene  CCC GCG CCG TCC TCC TCC TCC GGC TCC GGC TCC GCC CGC CCC AGG CCC AGG CTC ATC
gi606814  CCC GCG CCG TCC TCC TCC TCC GGC TCC GGC --- --- CGC CCC AGG CTC AGG CTC ATC
gi606810  CCC GCG CCG TCC TCC TCC TCC GGC TCC GGC --- --- CGC CCC AGG CTC AGG CTC ATC
           P   A   P   S   S   S   S   G   S   G           R   P   R   L   R   L   I

AZM4_68974  CGG AAC GCC CCC GTC TTC GCC GCC CCC GCC ACC GTC GT
CA2 Gene    CGG AAC GCC CCC GTC TTC GCC GCC CCC GCC ACC GTC GT
gi606814    CGG AAC G-C CCC GTC TTC GCC GCC CCC GCC ACC GTC GT
gi606810    CGG AAC GCC CCC GTC TTC GCC GCC CCC GCC ACC GTC GT
             R   N   A   P   V   F   A   A   P   A   T   V  (V)
```

C. ClustalW alignment of Intron 1 of AZM4_68974 with the CA2 gene.

```
AZM4_68974      1 GTACGTACGTGCGTACGGAGTACGACAATTAATGCATGCATGGCTCACTG    50
CA2 Gene        1 GTACGTACGTGCGTACGGTGTACGACAATTAATGCATGCATGGCTCACTG    50
                  ****************** *******************************

AZM4_68974     51 CACAGCGAGCACATCATGCACTGTACAGCATGCATGTCCTCGTTTCTATA   100
CA2 Gene       51 CACAGCGAGGACATCATGCACTGTACAGCATGCATGTCCTCGTTTCTATA   100
                  ********* ****************************************

AZM4_68974    101 TATTCTATCCACGTACGCCTTCGCTTCATCCAGCTCTAATAACCAAGTAC   150
CA2 Gene      101 TATTCTATCCACGTACGCCTTCGCTTCATCCAGCTCTAATAACCAAGTAC   150
                  **************************************************

AZM4_68974    151 TGACTAGTCGGCACTACTGACGACCTTGCTGTTTTGGGCACGAAATGCAC   200
CA2 Gene      151 TGACTAGTCGGCACTACTGACGACCTTGCTGTTTTGGGCACGAAATGCAC   200
                  **************************************************

AZM4_68974    201 TCATGCAACCAAAAGTCCTTGCTCCCATTATCTAGGAGGACACGACCAAG   250
```

```
CA2 Gene         201 TCATGCAACCAAAAGTCCTTGCTCCCATTATCTAGGAGGACACGACCAAG  250
                     **************************************************

AZM4_68974       251 CATGCAGGATATTCT--GGAAACCCAATCAAAATCCGACTG-TAATATAA  297
CA2 Gene         251 CATGCAGGATATTCCTGGAAACCCCAATCAAAATCCGACTGGTAATATAA  300
                     **************   * ** ******************* ********

AZM4_68974       298 GTATAATGCACATCCC-AAGAAGACGCCACTCTTAGCTTCATCTTAAATT  346
CA2 Gene         301 GTATAATGCACATCCCCAAGAAGACGCCACTCTTAGCTTCATCTTTAAAT  350
                     **************** **************************** ** *

AZM4_68974       347 ---CCCTTAAATTTAGCTATATATCA-TTTATATAACATC--AAAACAGG  390
CA2 Gene         351 TCCCCCTTAAATTTAGCTATATATCAATTTATTTAACAATCAAAAACAGG  400
                        *********************** ***** *****    ********

AZM4_68974       391 CGTACTATCAAAAATATATTTCATGGTACGTAAATATGGCACACTACACT  440
CA2 Gene         401 CGTACTATCAAAAATTTTTTTTATGGTACGTAAATATGGCACACTACACT  450
                     *************** * *** ****************************

AZM4_68974       441 GTACATAAATTTTTGTTTGAATTAATTCTTTCTGTCAAATCTATAATCAA  490
CA2 Gene         451 GTACATAAATTTTTGTTTGAATTAATTCTTTCTGTCAAATCTATAATCAA  500
                     **************************************************

AZM4_68974       491 ATTCAAGGCAGTTTGATATATACGTTAGATCTATATATAATGCATCTATT  540
CA2 Gene         501 ATTCAAGGCAGTTTGATATATACGTTAGATCTATATATAATGCATCTATT  550
                     **************************************************

AZM4_68974       541 TCTGACGGAGGGAATAGCTAGTGATGATGATAGTAATTTAGATCATTTTC  590
CA2 Gene         551 TCTGGCGGAGGGAATAGCTAGTGATGATGATAGTAATTTAGATCATTTTC  600
                     **** *********************************************

AZM4_68974       591 CCATCAGCTAGTAGCTACCGACGACATACGCATGTCAGCCATCTCCAATA  640
CA2 Gene         601 CCATCAGCTAGTAGCTACCGACGACATACGCATGTCAGCCATCTCCAATA  650
                     **************************************************

AZM4_68974       641 GAATATTCCCGAAGGGAGGTGTTTCCAAAAAGATGACGGCCAA-CGATAG  689
CA2 Gene         651 GAATATTCCCGAAGGGAGGTGTTTCCAAAAAGATGACGGGCAAACGATAG  700
                     *************************************** *** ******

AZM4_68974       690 TGCTAGTTTGAAGGCAACTACTACGTATATATCCTTTCAGTATAACAGAA  739
CA2 Gene         701 TGCTAGTTTGAAGGCAACTACTACGTATATATCCTTTCAGTATAACAGAA  750
                     **************************************************

AZM4_68974       740 TTCCACCCAGAAAAAAAAAGTCTCGAGTTGAATGAAAGAGGAGTAGTGAC  789
CA2 Gene         751 TTCCACCCAGAAAAAAAAAGTCTCGAGTTGAATGAAAGAGGAGTAGTGAC  800
                     **************************************************

AZM4_68974       790 GTCGAGCGCGCGTGAAATAAAGTATATGGCTGGCTTTTTCCTAAAGCGAT  839
CA2 Gene         801 GTCGAGCGCGCGTGAAATAAAGTATATGGCTGGCTTTTTCCTAAAGCGAT  850
                     **************************************************

AZM4_68974       840 AAGACCAGTTTATGCAGTGGGGTCATGGACATGTGTAGTGATAGCTAATA  889
CA2 Gene         851 AAGACCAGTTTATGCAGTGGGGTCATGGACATGTGTAGTGATAGCTAATA  900
                     **************************************************

AZM4_68974       890 ATCGTCCGCGTCTTTTGGCTTTTGAGTTCCGTTTGATCCATGACGCATAT  939
CA2 Gene         901 ATCGTCCGCGTCTTTTGGCTTTTGAGTTCCGTTTGATCCATGACGCATAT  950
                     **************************************************

AZM4_68974       940 ATATCCAGGCAGTTGAATAACCGACGACCATCAAATAAAAGGC-GCCACT  988
CA2 Gene         951 ATATCCAGGCAGTTGAATAACCGACGACCATCAAATAAAAGGCCGCCACT 1000
                     ******************************************* ******

AZM4_68974       989 ACTAGTGGCCATCGACGTCAGTTTAACCTTTCTATGTATGCATGTGTAAC 1038
CA2 Gene        1001 ACTAGTGGCCATCGACGTCAGTTTAACCTTTCTATGTATGCATGTGTAAC 1050
                     **************************************************

AZM4_68974      1039 TTCCCATGATTTCCTGCGTCGCGTTATTTGCTTTGTTTCACCGTCGGAC 1088
CA2 Gene        1051 TTCCCATGATTTCCTTGGCTGCGTTATTTGCTTTGTTTCACCGTCGGAC 1100
                     ***************  *  ******************************

AZM4_68974      1089 GACGAAGTCTTTTAGATAGCAATAAGGAACTATATCTAAGTGCTAGTTTG 1138
CA2 Gene        1101 GACGAAGTCTTTTAGATAGCAATAAGGAACTATATCTAAGTCCTAGTTTG 1150
                     ***************************************** ********

AZM4_68974      1139 GGAACCTCGTTTTCCCACGAGATTTTCATTTTCCTAAGGTAAATTAGTTC 1188
CA2 Gene        1151 GGAACCTCGTTTTCCCACGAGATTTTCATTTTCCTAAGGTAAATTAGTTC 1200
                     **************************************************
```

```
AZM4_68974   1189 ATTTTTTTTTGAAAATAAGAATCTTTTGAAAAAGATG-TAATTATCAAAC 1237
CA2 Gene     1201 CGGCTTTTTTGAAAATAAGAATCTTTTGAAAAAGATGGTAATTATCAAAC 1250
                      ********************************* ************

AZM4_68974   1238 TAGTCCTAACAGAGAGATTTTTGAGGGGGGAGAAAAAAAAGGAAGTTCTT 1287
CA2 Gene     1251 TAGTCCTAACAGAGAGATTTTTGAGGGGGGAGAAAAAAAAGGAAGTTCTT 1300
                  **************************************************

AZM4_68974   1288 CTGCATTCTTTTTTGGAGGAACAAAAAATTTGCCTCTGCATACTGAATCA 1337
CA2 Gene     1301 CTGCATTCTTTTTTGGAGGAACAAAAAATTTGCCTCTGCATACTGAATCA 1350
                  **************************************************

AZM4_68974   1338 GAGGGGATGGGCTTTATTTCGTGTTGGCTGGTTGATTGATGATTGGATGA 1387
CA2 Gene     1351 GAGGGGATGGGCTTTATTTCGTGTTGGCTGGTTGATTGATGATTGGATGA 1400
                  **************************************************

AZM4_68974   1388 GCTCCAGTAAGTTTGGAAGAGAACAGGGCACGGTCCCGACGGTTGGTACG 1437
CA2 Gene     1401 GCTCCAGTAAGTTTGGAAGAGAACAGGGCACGGTCCCGACGGTTGGTACG 1450
                  **************************************************

AZM4_68974   1438 GGTGAAGAAAGGGAGTGATTTAATTTATCGCCCCAACCACAACCACCCAT 1487
CA2 Gene     1451 GGTGAAGAAAGGGAGTGATTTAATTTATCGCCC-AACCACAACCACCCAT 1499
                  ********************************* ****************

AZM4_68974   1488 CGATCTATAGTTGCAGAAGAACTCGCTAATCCTGTCCACAAAAGCCGCAC 1537
CA2 Gene     1500 CGATCTATAGTTGCAGAAGAACTCGCTAATGCTGTCCACAAAAGCCGCAC 1549
                  ****************************** *******************

AZM4_68974   1538 TCACGCACTCATCCGCCACTGATTTTATTTCCCCCCCCCCCCCCTGTGGG 1587
CA2 Gene     1550 TCACGCACTCATCCGCCACTGATTTTATTTCCCCCCCCCCCCCT----GT 1595
                  *******************************************     *

AZM4_68974   1588 CGCGCGCGCGTGCTGCGTGGTGGTACTACTACCTGTTTGT-CTCACTGAC 1636
CA2 Gene     1596 GGCGCGCGGTTGCTGCGTGGTGGTACTACTACCTGTTTTTGCTCACTGAC 1645
                   *******  **************************** * *********

AZM4_68974   1637 ACAGTTGCGCGCGTCATCATGTTGCT 1662
CA2 Gene     1646 ACAGTTGCGGGT-TCATCATGTTGCT 1670
                  ********* *  *************
```

D. Alignment of AZM4_68974 and the CA2 gene with Exon 2 of the CA2 cDNA sequence (Burnell and Ludwig, 1997; gi:606810). Differences in the sequences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68974    AGT AAA CGG GAC GGC GGG CAG CTG AGG AGT CAA ACG AGA GAG ATC GAG AGA --A AGA
CA2 Gene   AGT AAA CGG GAC GGC GGG CAG CTG AGG AGT CAA ACG AGA GAG ATC GAG AGA GAA AGA
gi606810   TGT AAA CGG GAC GGC GGG CAG CTG AGG AGT CAA ACG AGA GAG ATC GAG AGA GAA AGA
            C   K   R   D   G   G   Q   L   R   S   Q   T   R   E   I   E   R   E   R

4_68974    AAG GGA GGG CAT CCA CCA GCC GCC GGC GAT AAG AGG GGA GGA GAG AGA GGC CAG AGA
CA2 Gene   AAG GGA GGG CAT CCA CCA GCC GGC GGG CAT AAG AGG GGA GGA GAG AGA GGC CAG AGA
gi606810   AAG GGA GGG CAT CCA CCA GCC GGC GGG CAT AAG AGG GGA GGA GAG AGA GGC CAG AGA
            K   G   G   H   P   P   A   G   G   H   K   R   G   G   E   R   G   G   R

4_68974    AGA GGA GGA GAA GAA GAA GAA AAT GAG CAG CTG CCT CTG CCT TCC GAA AAA AAA GGA
CA2 Gene   AGA GGA GGA GAA GAA GAA GAA GAT GAG CAG CTG CCT CTG CCT TCC GAA AAA AAA GGA
gi606810   AGA GGA GGA GAA GAA GAA GAA GAT GAG CAG CTG CCT CTG CCT TCC GAA AAA AAA GGA
            R   G   G   E   E   E   E   D   E   Q   L   P   L   P   S   E   K   K   G

4_68974    GGG GCC AGC GAA GGA GAA GCC GTC CAC AGA TAC CCC CAC CTC GTC ACT CCT TCA GAA
CA2 Gene   GGG GCC AGC GAA GGA GAA GCC GTC CAC AGA TAC CCC CAC CTC GTC ACT CCT TCA GAA
gi606810   GGG GCC AGC GAA GGA GAA GCC GTC CAC AGA TAC CCC CAC CTC GTC ACT CCT TCA GAA
            G   A   S   E   G   E   A   V   H   R   Y   P   H   L   V   T   P   S   E

4_68974    CCA GAA GCC CTC CCA A CCT CCA CCT CCT CCC TCC AAG GCT TCC TCC AAG GTC
CA2 Gene   CCA GAA GCC CTC C-A A CCT CCA CCT CCT CCC TCC AAG GCT TCC TCC AAG GGC
gi606810   CCA GAA GCC CTC C-A A CCT CCA CCT CCT CCC TCC AAG GCT TCC TCC AAG GGC
```

```
            P   E   A   L   Q     P   P   P   P   P   S   K   A   S   S   K   G
```

E. ClustalW alignment of Intron 2 of AZM4_68974 with the CA2 gene.

```
AZM4_68974       1 CGTCCCCTCCTCCTCCTCCTCATCTTCCTCTCTCACCTTCAGCACCATCC 50
CA2 Gene         1 CGTCCCCTCCTCCTCCTCCTCATCTTCCTCTCTCACCTTCAGCACCATCC 50
                   **************************************************

AZM4_68974      51 TCCACACAGCAGCACGCGCGCAGCAATCTCACCGTTTTCTTTTCCTCCAT 100
CA2 Gene        51 TCCACACAGCAGCACGCGCGCAGCAATCTCACCGTTTTCTTTTCCTCCAT 100
                   **************************************************

AZM4_68974     101 TGCCATCAGTAGCTAGCCACACTGCATGCATTCAGCTTCCGCTTTCTCCC 150
CA2 Gene       101 TGCCATCAGTAGCTAGCCACACTGCATGCATTCAGCTTCCGCTTTCTCCC 150
                   **************************************************

AZM4_68974     151 TGTGTAGCGAGCGCTG-TGCCGGCCGGTGCAGAG                 183
CA2 Gene       151 TGTGTAGCGAGCGCTGGTGCCGGCCGGTGCAGAGAAGATCCCTGCTCCCC 200
                   **************** *****************

CA2 Gene       201 CCCCCCCCCCCCCCCCCTAATTAAGATCACCTTTGTGCATTTTTTTCCTT 250

CA2 Gene       251 GTGTTGTGGTCCGTCGGCAAGTAGGCCAAAATTGCATCATGCCATGGCCC 300

CA2 Gene       301 CTCCTCTTCTACTACCTCGTCATGCAGCCAGCAACGACATGAATGACCCG 350

CA2 Gene       351 AACGAAGTATCTGGCGTTGACATTGCAG 378
```

3.5: Alignments of the nucleotide sequence of the genomic DNA library clone with the AZM4_68973 assembly.

This alignment was made between the AZM4_68973 assembly and the second assembly of 2.7 kb obtained from screening the maize genomic DNA library (Fig. 3.13).

A. Alignment of AZM4_68973 and the genomic DNA library clone with Exon 11 of the CA2 cDNA sequence (Burnell and Ludwig, 1997; gi:606810). Differences in the sequences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68973   TAC ATG GTG TTC GCT TGC TCC GAC TCC CGT GTG TGC CCA TCG GTG ACC CTG GGC CTG
Glib        C ATG GTG TTC GCC TGC TCC GAC TCC CGT GTG TGC CCA TCG GTG ACC CTG GGC CTG
gi:606810 TAC ATG GTG TTC GCC TGC TCC GAC TCC CGT GTG TGC CCG TCG GTG ACC CTG GGC CTG
           Y   M   V   F   A   C   S   D   S   R   V   C   P   S   V   T   L   G   L

4_68973   CAG CCC GGC GAG GCC TTC ACC GTT CGC AAC ATA GCC GCC ATG GTC CCA GGC TAC GAC
Glib      CAG CCC GGC GAG GCC TTC ACC GTT CGC AAC ATA GCC GCC ATG GTC CCA GGC TAC GAC
gi:606810 CAG CCC GGC GAG GCC TTC ACC GTT CGC AAC ATC GCC GCC ATG GTC CCC GGC TAC GAC
           Q   P   G   E   A   F   T   V   R   N   I   A   A   M   V   P   G   Y   D

4_68973      AAG
Glib Clone1  AAG
gi:606810    AAG
              K
```

B. AZM4_68973 sequence of Intron 11 aligned with the genomic DNA library clone.

```
AZM4_68974         1 GTATATATACACACTGACGATTGTGAACAACGCAATGGTCTCAATTTCTA   50
Gen lib Clone1     1 GTATATATACACACTGACGATTGTGAACAACGCAATGGTCTCAATTTCTA   50
                     **************************************************

AZM4_68973        51 CTCACACGGCCGGCCGCGGCCTCTCGTTTTCGTGTCGACTGCAG  94
Gen lib Clone1    51 CTCACACGGCCGGCCGCGGCCTCTCGTTTTCGTGTCGACTGCAG  94
                     ********************************************
```

C. Alignment of AZM4_68973 and Exon 12 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810) with the genomic DNA library clone. Differences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68973   ACC AAG TAC ACC GGC ATC GGG TCC GCC ATC GAG TAC GCT GTG TGC GCT CTC AAG GTG
Glib      ACC AAG TAC ACC GGC ATC GGG TCC GCC ATC GAG TAC GCT GTG TGC GCT CTC AAG GTG
gi:606810 ACC AAG TAC ACC GGC ATC GGG TCC GCC ATC GAG TAC GCT GTG TGC GCC CTC AAG GTG
           T   K   Y   T   G   I   G   S   A   I   E   Y   A   V   C   A   L   K   V

4_68973   GAG GTC CTC GTG GTC ATT GGC CAT AGC TGC TGC GGT GGC ATC AGG GCG CTC CTC TCC
Glib      GAG GTC CTC GTG GTC ATT GGC CAT AGC TGC TGC GGT GGC ATC AGG GCG CTC CTC TCA
gi:606810 GAG GTC CTC GTG GTC ATT GGC CAT AGC TGC TGC GGT GGC ATC AGG GCG CTC CTC TCA
           E   V   L   V   V   I   G   H   S   C   C   G   G   I   R   A   L   L   S


AZM4_68973   CTC CAG GAC GGC GCA CCT GAC ACC TT
Glib         CTC CAG GAC GGC GCA CCT GAC ACC TT
gi:606810    CTC CAG GAC GGC GCA CCT GAC ACC TT
              L   K   D   G   A   P   D   N  (F)
```

D. AZM4_68973 sequence of Intron 12 aligned with the genomic DNA library clone.

```
AZM4_68974         1 GTAAGTCGCGACAGTAAAATATATACAAGTTTCATTTA----GATATAAA   46
Gen lib Clone1     1 GTAAGTCGCCACCGTAAAATATA--CAAGTTTCATTTAATTAGATATAAA   48
                     ********* ** **********  *************    ********

AZM4_68973        47 AAA-CTATTTGCGCTTATTTATGTCATGCATGATTTTGATCCTCTCTATA   95
Gen lib Clone1    49 AAAATTATTTGCGCTTATTTATGTCATGCATGATTTTGATCCTCTCTA--   96
                     ***  *******************************************

AZM4_68973        96 CCATGTTGTGTGTTGGTTTGGTGTGGTGTACGTACGCAG  134
Gen lib Clone1    97 CCATGTTGTGTGTTGGTTTGGTGTGGTGTACGTACGCAG  135
                     ***************************************
```

E. Alignment of AZM4_68973 and Exon 13 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810) with the genomic DNA library clone. Differences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68973   --C CAC TTC GTC GAG GAC TGG GTT AAG ATC GGC TTC ATT GCC AAG ATG AAG GTA AAG
Glib      --C CAC TTC GTC GAG GAC TGG GTT AAG ATC GGC TTC ATT GCC AAG ATG AAG GTA AAG
gi:606810 --C CAC TTC GTC GAG GAC TGG GTT AAG ATC GGC TTC ATT GCC AAG ATG AAG GTA AAG
               H   F   V   E   D   W   V   K   I   G   F   I   A   K   N   K   V   K

4_68973      AAA GAG CAC GCC TCG GTG CCG TTC GAT GAC CAG TGC TCC ATT CTC GAG AAG
Glib         AAA GAG CAC GCC TCG GTG CCG TTC GAT GAC CAG TGC TCC ATT CTC GAG AAG
gi:606810    AAA GAG CAC GCC TCG GTG CCG TTC GAT GAC CAG TGG TCC ATT CTC GAG AAG
```

```
               K   E   H   A   S   V   P   F   D   D   Q   C   S   I   L   E   K
```

F. AZM4_68973 sequence of Intron 13 aligned with the genomic DNA library clone.

```
AZM4_68973         1 GTATGTTGTACATTCGTCGAGCAGTTACTGTTGCATGAATAGATTGGTTT 50
Gen lib Clone1     1 GTATGTTGTACATTCGTCGAGCAGTTACTGTTGCATGAATAAATTGTTTT 50
                     **************************************** **** ***

AZM4_68973        51 TTGCTCACCAAAAGGACCTCTATTGTTTCTGCAG   84
Gen lib Clone1    51 TTGCTCACCA-AAGGACCTCTATTGTTTCTGCAG   83
                     ********** ***********************
```

G. Alignment of AZM4_68973, the genomic DNA library clone and the last exon of the CA cDNA sequence, Exon 14 (Burnell and Ludwig, 1997; gi:606810 and gi:606814). This exon corresponds specifically to Repeat C of these cDNA sequences, and the last 221 bp represent the 3′-untranslated region. Differences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68973   GAG GCC GTG AAC GTG TCC CTG GAG AAC CTC AAG ACC TAC CCC TTC GTC AAG GAA GGG
Glib      GAG GCC GTG AAC GTG TCC CTG GAG AAC CTC AAG ACC TAC CCC TTC GTC AAG GAA GGG
606810/4  GAG GCC GTG AAC GTG TCC CTG GAG AAC CTC AAG ACC TAC CCC TTC GTC AAG GAA GGG
           E   A   V   N   V   S   L   E   N   L   K   T   Y   P   F   V   K   E   G


4_68973   CTT GCA AAT GGG ACC CTC AAG CTG ATC GGC GCC CAC TAC GAC TTT GTC TCA GGA GAG
Glib      CTT GCA AAT GGG ACC CTC AAG CTG ATC GGC GCC CAC TAC GAC TTT GTC TCA GGA GAG
606810/4  CTT GCA AAT GGG ACC CTC AAG CTG ATC GGC GCC CAC TAC GAC TTT GTC TCA GGA GAG
           L   A   N   G   T   L   K   L   I   G   A   H   Y   D   F   V   S   G   E

4_68973   TTC CTC ACA TGG AAA AAG TGA AAA ACT AGG GCT ACG GCA ATT CTA CCG GCC CGC CGA
Glib      TTC CTC ACA TGG AAA AAG TGA AAA ACT AGG GCT AAG GCA ATT CTA CCG GCC CGC CGA
606810/4  TTC CTC ACA TGG AAA AAG TGA AAA ACT AGG GCT AAG GCA ATT CTA CCG GCC CGC CGA
           F   L   T   W   K   K   *

4_68973   CTC CTG CAT CAT CAT AAA TAT ATA TAC T-A TA- CTA TAC TAC TAC GTA CCT ACC GAT
Glib      CTC CTG CAT CAT CAT AAA TAT ATA TAC T-A TAA CTA TAC TAC TAC GTA CCT ACC GAT
606810/4  CTC CTG CAT CAT CAT AAA TAT ATA TAC TCT AA- CTA TAC TAC TAC GTA CCT ACC GAT

4_68973   ATG CAC CCG AGC AAT GTG AAT GCG TCG AGT ACT ATA TAT CTG TTT TCT GCA TCT ACA
Glib      ATG CAC CCG AGC AAT GTG AAT GCG TCG AGT ACT --- -AT CTG TTT TCT GCA TCT ACA
606810/4  ATG CAC CCG AGC AAT GTG AAT GCG TCG AGT ACT --- -AT CTG TTT TCT GCA TCT ACA

4_68973   TAT ATA TAC CGG ATC AAT CGC CCA ATG --- --- --T GAA TGT AAT AAG CAA TAT CAT
Glib      TAT ATA TAC CGG ATC AA- --- -CA ATC GCC CAA TGT GAA TGT AAT AAG CAA TAT CAT
gi:606810 TAT ATA TAC CGG ATC AAT CGC CCA ATG --- --- --T GAA TGT AAT AAG CAA TAT CAT

4_68973      TTT CTA CCA CTT TTC ATT CCT AA
Glib         TTT CTA CCA CTT TTC ATT CCT AA
gi:606810    TTT CTA CCA CTT TTC ATT CCT AA
```

3.6 Nucleotide sequence alignment of AZM4_68973 and the CA2 gene.

A. Alignment of AZM4_68973 and the CA2 gene with Exon 3 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810 and gi:616814). Differences in the sequences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68973    GGC ATG GAC CCC ACC GTC GAG CGC TTGGAAG AGC GGG TTC CAG AAG TTC AAG ACC GAG
CA2 Gene   GGC ATG GAC CCC ACC GGC GAG CGC TTG-AAG AGC GGG TTC CAG AAG TTC AAG ACC GAG
606810/4   GGC ATG GAC CCC ACC GTC GAG CGC TTG-AAG AGC GGG TTC CAG AAG TTC AAG ACC GAG
            G   M   D   P   T   V   E   R   L   K   S   G   F   Q   K   F   K   T   E

4_68973       GTC TAT GA
CA2 Gene      GTC TAT GA
606810/4      GTC TAT GA
               V   Y  (D)
```

B. ClustalW alignment of Intron 3 of AZM4_68973 with the CA2 gene.

```
AZM4_68973      1 GTAAGTCACCTGAGCTGTTTGTTCTCTGCAGCA-CCCGCGTTTGGTTTCT 49
CA2 Gene        1 GTAAGTCACCTGAGCTGTTTGTTCTCTGCAGCAACCCGCGTTTGGTTTCT 50
                  ********************************* ****************

AZM4_68973     50 ATTTCCTTTTTTGTTTGTTTGTGAATTCAGTGAGCTCCGACTCCGACTGA 99
CA2 Gene       51 ATTTCCTTTTTT-TTTGTTTGTGAATTCAGTGAGCTCCGACTCCGACTGA 99
                  ************ *************************************

AZM4_68973    100 TCATGTGCTCCGCTGATCTTTGTTCGCAG 129
CA2 Gene      100 TCATGTGCTCCGCTGATCTTTGTTCGCAG 129
                  *****************************
```

C. Alignment of AZM4_68973 and the CA2 gene with Exon 4 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810 and gi:616814). There are no differences in the three sequences. The consensus translated sequence is also shown.

```
AZM4_68973    --C AAG AAG CCG GAG CTG TTC GAG CCT CTC AAG TCC GGC CAG AGC CCC AG-
CA2 Gene      --C AAG AAG CCG GAG CTG TTC GAG CCT CTC AAG TCC GGC CAG AGC CCC AG-
gi:606810/4   --C AAG AAG CCG GAG CTG TTC GAG CCT CTC AAG TCC GGC CAG AGC CCC AG-
                   K   K   P   E   L   F   E   P   L   K   S   G   Q   S   P   R
```

D. ClustalW alignment of Intron 4 of AZM4_68973 with the CA2 gene.

```
AZM4_68973      1 GTATGCGCTGCTAATGTTTTTTTATATATATTTTGTTGTGTGTCTATAGC 50
CA2 Gene        1 GTATGCGCTGCTAATGTTTTTTTATATATATTTTGTTGTGTGTCTATAGC 50
                  **************************************************

AZM4_68973     51 GACTCCGGCCAACTGGGCCAAAAGATTGA-GTAGTACTAGTTGCTCGTTC 99
CA2 Gene       51 GACTCCGGGCAACTGGGGCAAAAGATTGAAGTAGTACTAGTTGCTCGTTC 100
                  ******** ******** *********** ********************

AZM4_68973    100 CTATTACTAGCTCTGTAGCTCATCACCATTGCTGCTGCAACACCCTGCCG 149
CA2 Gene      101 CTATTACTAGCTCTGTAGCTCATCACCATTGCTGCTGCAACACCCTGCCG 150
                  **************************************************


AZM4_68973    150 CACCTGCACTATTCAGCATCCACCCTGTCTCCCCTGGACCAAAGCTGCAA 199
CA2 Gene      151 CACCTGCACTATTCAGCATCCACCCTGTCTCCCCTGGACCAAAGCTGCAA 200
                  **************************************************
```

```
AZM4_68973    200 GGGGAACCATGCAGATAATACTAGGTGTGTATTATCAGCATTCCATGGCT 249
CA2 Gene      201 GGGGAACCATGCAGATAATACTAGGTGTGTATTATCAGCATTCCATGGCT 250
                  **************************************************

AZM4_68973    250 AATGTGTGGTCCAGGCGTCCAGCACTGTCCCTCGCCCCACCTCACGGGAT 299
CA2 Gene      251 AATGTGTGGTCCAGGCGTCCAGCACTGTGGGTCGCCCCAC-TCACGGGAT 299
                  ****************************   ********* *********

AZM4_68973    300 CCTGTCGTCATCGTGAGTAGTTGGCTTGGACGTGTCCCCTTCCCCTCTCG 349
CA2 Gene      300 CCTGTCGTCATCGTGAGTAGTTGGCTTGGACGTGTCCCCTTCCCCTCTCG 349
                  **************************************************

AZM4_68973    350 CACCCCTTGCAAAAAAGTTAGGTGCATAAATGTTGGGCCTGTTGCCGGTC 399
CA2 Gene      350 CACCCCTTGCAAAAAAGTTAGGTGCATAAATGTTGGGCCTGTTGCCGGTC 399
                  **************************************************

AZM4_68973    400 CTCGAGGAAATATGCTACACTACAGATGTCCCAATTTTTGTGGAAGATAT 449
CA2 Gene      400 CTCGAGGAAATATGCTACAC-ACAGATGTCCCAATTTTTGTGGAAGATAT 448
                  ******************** *****************************

AZM4_68973    450 GGCAGCAGCATCACGCCTCCTGATGATGCCCGGAACGGAAATGTTCTTGC 499
CA2 Gene      449 GGCAGCAGCATCACGCCTCCTGATGATGCCCGGAACGGAAATGTTCTTGC 498
                  **************************************************

AZM4_68973    500 TATTGGCCGCCAGCAGGG-AATATAATGGGATAAAGATAGACCAGCGTGC 548
CA2 Gene      499 TATTGGCCGCCAGCAGGGGAATATAATGGGATAAAGATAGACCAGCGTGC 548
                  ****************** *******************************

AZM4_68973    549 TAGAGAGCCACACGGAAACCAGAGCGCGCGTAGAGCATCCTCGTCGCAAC 598
CA2 Gene      549 TAGAGAGCCACACGGAAACCAGAGCGCGCGTAGAGCATCCTCGTCGCAAC 598
                  **************************************************

AZM4_68973    599 TAATACTAGTACTTACAGAGCCAGAGGAGGAGGGTCAAATCGAAACTCAA 648
CA2 Gene      599 TAATACTAGTACTTACAGAGCCAGAGGAGGAGGGTCAAATCGAAACTCAA 648
                  **************************************************

AZM4_68973    649 TCAAAAGCTTGCCGCCTTTTTGGGGCGCCAGAAATCTTCCACTGATGAGA 698
CA2 Gene      649 TCAAAAGCTTGTCGCCTTTTTGGGGCGCCAGAAATCTTCCACTGATGAGA 698
                  *********** **************************************

AZM4_68973    699 TGACCAGGGCCGATGATCTGCTTACCTGCTTATCGATAAGAGCCATGGGA 748
CA2 Gene      699 TGACCAGGGCCGATGATCTGCTTACCTGCTTATCGATAAGAGCCATGGGA 748
                  **************************************************

AZM4_68973    749 AACCGATCGAACTTGGTTTTGCGTACGTGCTCCTCCCTCTTTTCACCGAC 848
CA2 Gene      749 AACCGATCGAACTTGGTTTTGCGTACGTGCTCCTCCCTCTTTTCACCGAC 848
                  **************************************************

AZM4_68973    849 CTGACGGTGACTGATTTCCCCTCCGCTGCAG 880
CA2 Gene      849 -TGACGGTGACTGATTTCCCCTCCGCTGCAG 879
                   *******************************
```

E. Alignment of AZM4_68973 and the CA2 gene with Exon 5 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810 and gi:616814). There are no differences in the three sequences. The consensus translated sequence is also shown.

```
4_68973    --G TAC ATG GTG TTC GCC TGC TCC GAC TCC CGC GTG TGC CCG TCG GTG ACC CTG GGA
CA2 Gene   --G TAC ATG GTG TTC GCC TGC TCC GAC TCC CGC GTG TGC CCG TCG GTG ACC CTG GGA
606810/4   --G TAC ATG GTG TTC GCC TGC TCC GAC TCC CGC GTG TGC CCG TCG GTG ACC CTG GGA
                Y   M   V   F   A   C   S   D   S   R   V   C   P   S   V   T   L   G

4_68973    CTG CAG CCC GGC GAG GCA TTC ACC GTC CGC AAC ATC GCT TCC ATG GTC CCA CCC TAC
CA2 Gene   CTG CAG CCC GGC GAG GCA TTC ACC GTC CGC AAC ATC GCT TCC ATG GTC CCA CCC TAC
606810/4   CTG CAG CCC GGC GAG GCA TTC ACC GTC CGC AAC ATC GCT TCC ATG GTC CCA CCC TAC
            L   Q   P   G   E   A   F   T   V   R   N   I   A   S   M   V   P   P   Y

4_68973        GAC AAG
CA2 Gene       GAC AAG
606810/4       GAC AAG
                D   K
```

F. ClustalW alignment of Intron 5 of AZM4_68973 with the CA2 gene.

```
AZM4_68973       1 GTACGTACGAGCAAACACCGATCGACGCATGCAACGGTGGTATCAGCCAC 50
CA2 Gene         1 GTACGTACGAGCAAACACCGATCGACGCATGCAACGGTGGTATCAGCCAC 50
                   **************************************************

AZM4_68973      51 ACTAATATTACTCACACGGTCGTCTTCCGTTTTGGCCAAACTGCAG 96
CA2 Gene        51 ACTAATATTACTCACACGGTCGTCTTCCGTTTTGGCCAAACTGCAG 96
                   **********************************************
```

G. Alignment of AZM4_68973 and the CA2 gene with Exon 6 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810 and gi:616814). Differences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68973   ATC AAG TAC GCC GGC ACC GGG TCC GCC ATC GAG TAC GCC GTG TGC GCG CTC AAG GTG
CA2 Gene  ATC AAG TAC GCC GGC ACC GGG TCC GCC ATC GAG TAC GCC GTG TGC GCG CTC AAG GTG
606810/4  ATC AAG TAC GCC GGC ACA GGG TCC GCC ATC GAG TAC GCC GTG TGC GCG CTC AAG GTG
           I   K   Y   A   G   T   G   S   A   I   E   Y   A   V   C   A   L   K   V

4_68973   CAG GTC ATC GTG GTC ATT GGC CAC AGC TGC TGC GGT GGC ATC AGG GCG CTC CTC TCC
CA2 Gene  CAG GTC ATC GTG GTC ATT GGC CAC AGC TGC TGC GGT GGC ATC AGG GCG CTC CTC TCC
606810/4  CAG GTC ATC GTG GTC ATT GGC CAC AGC TGC TGC GGT GGC ATC AGG GCG CTC CTC TCC
           Q   V   I   V   V   I   G   H   S   C   C   G   G   I   R   A   L   L   S

4_68973      CTC AAG GAC GGC GCG CCC GAC AAC TT
CA2 Gene     CTC AAG GAC GGC GCG CCC GAC AAC TT
606810/4     CTC AAG GAC GGC GCG CCC GAC AAC TT
              L   K   D   G   A   P   D   N  (F)
```

H. ClustalW alignment of Intron 6 of AZM4_68973 with the CA2 gene.

```
AZM4_68973       1 GTAAGCAGTAGTCATCGTAAAATGCGTATAAAAAATATATATAGCAGTTT 50
CA2 Gene         1 GTAAGCAGTAGTCATCGTAAAATGCGTATAAAAAATATATATAGCAGTTT 50
                   **************************************************

AZM4_68973      51 TATTTAGAGAGAGAGAAAAAAATTAGAACCCCGTGTAGTGTAACCTGCTC 100
CA2 Gene        51 TATTTAGAGAGAGAGAAAAAAATTAGAACCCCGTGTAGTGTAA--TGCTC 98
                   *******************************************  *****

AZM4_68973     101 AGCGTGTTGTCTGTCGTTGGTTTAAATCTGGCCATGTATATCCAG 145
CA2 Gene        99 AGCGTGTTGTCTGTCGTTGGTTTAAATCTGGCCATGTATATCCAG 143
                   *********************************************
```

I. Alignment of AZM4_68973 and the CA2 gene with Exon 7 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810 and gi:616814). Differences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68973   --C CCA TTC GTG GAG GAC TGG GTC AGG ATC GGC AGC CCT GCC AAG AAC AAG GTG AAG
CA2 Gene  --C CCA TTC GTG GAG GAC TGG GTC AGG ATC GGC AGC CCT GCC AAG AAC AAG GTG AAG
606810/4  --C CAC TTC GTG GAG GAC TGG GTC AGG ATC GGC AGC CCT GCC AAG AAC AAG GTG AAG
               H   F   V   E   D   W   V   R   I   G   S   P   A   K   N   K   V   K

4_68973      AAA GAG CAC GCA TCG GTG CCG TTC GAT GAC CAG TGC TCC ATC CTG GAG AAG
CA2 Gene     AAA GAG CAC GCA TCG GTG CCG TTC GAT GAC CAG TGC TCC ATC CTG GAG AAG
```

```
606810/4      AAA GAG CAC GCG TCC GTG CCG TTC GAT GAC CAG TGC TCC ATC CTG GAG AAG
               K   E   H   A   S   V   P   F   D   D   Q   C   S   I   L   E   K
```

J. ClustalW alignment of Intron 7 of AZM4_68973 with the CA2 gene.

```
AZM4_68973      1 GTACGTAACGTAAACGCACGCACACACACCGACCGTATGAATAATGGATT 50
CA2 Gene        1 GTACGTAACGTAAACGCACGCACACACACCGACCGTATGAATAATGGATT 50
                  **************************************************

AZM4_68973     51 ATATATTATTGGTTTCGCTCATCAACGAACAAATTCAAGGATCATCATCG 100
CA2 Gene       51 ATATATTATTGGTTTCGCTCATCAACGAACAAATTCAAGGATCATCATCG 100
                  **************************************************

AZM4_68973    101 ACCTTTAATTGTGTGTGTGTGTTTCTGCAG 130
CA2 Gene      101 ACCTTTAATTGTGTGTGTGTGTTTCTGCAG 130
                  ******************************
```

K. Alignment of AZM4_68973 and the CA2 gene with Exon 8 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810 and gi:616814). Differences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68973    GAG GCC GTG AAC GTG TCG CTC CAG AAC CTC AAG AGC TAC CCC TTC GTC AAG GAA GGG
CA2 Gene   GAG GCC GTC AAC GTC TCG CTC CAG AAC CTC AAG AGC TAC CCC TTC GTC AAG GAA GGG
606810/4   GAG GCC GTG AAC GTG TCG CTC CAG AAC CTC AAG AGC TAC CCC TTC GTC AAG GAA GGG
            E   A   V   N   V   S   L   Q   N   L   K   S   Y   P   F   V   K   E   G


4_68973    CTG GCC GGC GGG ACG CTC AAG CTG GTT GGC GCC CAC TAC GAC TTC GTC AAA GGG CAG
CA2 Gene   CTG GCC GGC GGG ACG CTC AAG CTG GTT GGC GCC CAC TAC GAC TTC GTC AAA GGG CAG
gi:606810 CTG GCC GGC GGG ACG CTC AAG CTG GTT GGC GCC CAC TCA GAC TTC GTC AAA GGG CAG
gi:606814 CTG GCC GGC GGG ACG CTC AAG CTG GTT GGC GCC CAC TAC AGC TTC GTC AAA GGG CAG
            L   A   G   G   T   L   K   L   V   G   A   H   Y   D   F   V   K   G   Q

4_68973       TTC GTC ACA TGG GAG CCT
CA2 Gene      TTC GTC ACA TGG GAG CCT
gi:606810/4   TTC GTC ACA TGG GAG CCT
               F   V   T   W   E   P
```

L. ClustalW alignment of Intron 8 of AZM4_68973 with the CA2 gene.

```
AZM4_68973      1 GTAGGGGTCCACGCGCACAGCTCTTCTTTTAGAGCAACTCCAATAGATTA 50
CA2 Gene        1 GTAGGGGTCCACGCGCACAGCTCTTCTTTTAGAGCAACTCCAATAGATTA 50
                  **************************************************

AZM4_68973     51 GCCAAATTTTTTATTCTATATTCTCATTTAGCTAGCCGTTTAGCTATAAT 100
CA2 Gene       51 GCCAAATTTTTTATTCTATATTCTCATTTAGCTAGCCGTTTAGCTATAAT 100
                  **************************************************

AZM4_68973    101 TCACTCTCTAAATTACGATTAATTCCAACAGACTAGCCAAATTAGACTGG 150
CA2 Gene      101 TCACTCTCTAAATTACGATTAATTCCAACAGACTAGCCAAATTAGACTGG 150
                  **************************************************

AZM4_68973    151 TAGGTCCCACATGTCATTCTCACCTTGCCTTCTTCCCTATGTCCCACGCG 200
CA2 Gene      151 TAGGTCCCACATGTCATTCTCACCTTGCCTTCTTCCCTATGTCCCACGCG 200
                  **************************************************

AZM4_68973    201 CCTATGCTACACCGTCCTGCCTCCACTCCGGCTGAGGACAAAGGCTATGG 250
CA2 Gene      201 CCTATGCTACACCGTCCTGCCTCCACTCCGGCTGAGGACAAAGGCTATGG 250
                  **************************************************

AZM4_68973    251 GAGGACCGGATAGCTAGCGTAGGGAATTAGTCATATTTGGCTAGCGGAGG 300
CA2 Gene      251 GAGGACCGGATAGCTAGCGTAGGGAATTAGTCATATTTGGCTAGCGGAGG 300
                  **************************************************

AZM4_68973    301 GGGTTGTTTACCGAGTTGGATAGCGAGAGAAGGATTTGAAGAGACTGTTG 350
```

```
CA2 Gene      301 GGGTTGTTTACCGAGTTGGATAGCGAGAGAAGGATTTGAAGAGACTGTTG 350
                  **************************************************

AZM4_68973    351 GATCCAATTTTTACTCCAATTTTATTATTTTTAGCTAGTCAATTTGTTTT 400
CA2 Gene      351 GATCCAATTTTTACTCCAATTTTATTATTTTTAGCTAGTCAATTTGTTTT 400
                  **************************************************

AZM4_68973    401 ACATAAGCTCTTGGAGTTGCTCTTACCTTTTTTTTTCAATTGCTATATTG 450
CA2 Gene      401 ACATAAGCTCTTGGAGTTGCTCTTACCTTTTTTTTTCAATTGCTATATTG 450
                  **************************************************

AZM4_68973    451 ACGACATCACGTCCGTCGTCTTGCATTTGCACATAGCTAGCGCACTCCAG 500
CA2 Gene      451 ACGACATCACGTCCGTCGTCTTGCATTTGCACATAGCTAGCGCACTCCAG 500
                  **************************************************

AZM4_68973    501 ATCCCAATTCCCAACATCATCCGGCCAGCCCCCTTTAATTTATCTCCCTT 550
CA2 Gene      501 ATCCCAATTCCCAACATCATCCGGCCAGCCCCCTTTAATTTATCTCCCTT 550
                  **************************************************

AZM4_68973    551 GTTTGCCATCGCAATTTCTTTCTCTCCCCTTAGCTTGTTGACATGCATGG 600
CA2 Gene      551 GTTTGCCATCGCAATTTCTTTCTCTCCCCTTAGCTTGTTGACATGCATGG 600
                  **************************************************

AZM4_68973    601 GAGGATATCAGGAGACGAAGAAAAGAGCAGAGCAGCGCCTTTGCCCTCCC 650
CA2 Gene      601 GAGGATATCAGGAGACGAAGAAAAGAGCAGAGCAGCGCCTTTGCCCTCCC 650
                  **************************************************

AZM4_68973    651 ATAGATTCCCACGCACCTCGTCACTCCTTGAGAACCAGAAGCCCACCCAC 700
CA2 Gene      651 ATAGATTCCCACGCACCTCGTCACTCCTTGAGAACCAGAAGCCCACCCAC 700
                  **************************************************

AZM4_68973    701 CCGGTCCAGTGTGGCCAAAAGTTGCATCACGCCCCCTTCATTCTCTCGCT 750
CA2 Gene      701 CCGGTCCAGTGTGCCCAAAAGTTGCATCACGCCCCCTTAATTCTCTCGCT 750
                  ************* *********************** ************

AZM4_68973    751 CTCTCTATACCCCCCTCATGCTGCATCTATCACCGTACCATCACGAGCAT 800
CA2 Gene      751 CTCTCAATACCCCCTTAATGCTGCATCTATCACCGTACCATCACGAGCAT 800
                  ***** ******** * *********************************

AZM4_68973    801 GCAAGTTAGTCTTTCCGGGCATGGCGAACTGACCGACGATTTTCTTGTTG 850
CA2 Gene      801 GCAAGTTAGTCTTTCCGGGNATGGCGAACTGACCGACGATTTTCTTGTTG 850
                  ******************* ******************************

AZM4_68973    851 CTGGTCCTGCAG 862
CA2 Gene      851 CTGGTCCTGCAG 862
                  ************
```

M. Alignment of AZM4_68973 and the CA2 gene with Exon 9 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810). Differences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68973    CCC CAG GAC GCC ATC GAG CGC TTG ACG AGC GGG TTC CAG CAG TTC AAG GTC AAT GTC
CA2 Gene   CCC CAG GAC GCC ATC GAG CGC TTG ACG AGC GGG TTC CAG CAG TTC AAG GTC AAT GTC
gi:606810  CCC CAG GAC GCC ATC GAG CGC TTG ACG AGC GGC TTC CAG CAG TTC AAG GTC AAT GTC
            P   Q   D   A   I   E   R   L   T   S   G   F   Q   Q   F   K   V   N   V

4_68973       TAT GA
CA2 Gene      TAT GA
gi:606810     TAT GA
               Y  (D)
```

N. ClustalW alignment of Intron 9 of AZM4_68973 with the CA2 gene. The CA2 gene sequence ends within this intron.

```
AZM4_68973      1 GTAAGTCACCCCTCTACTACTCAGAGCTGGCTGTTGTTTTCTG-CAGCAC 49
CA2 Gene        1 GTAAGTCACCCCTCTACTACTCAGAGCTGGCTGTTGTTTTCTGNCAGCAC 50
                  ******************************************* ******
```

```
AZM4_68973     50 CGGCGTTTGGTTTAGCCGTTTCAGTTTCAAACGTTTCATTTGTATCGGGA  99
CA2 Gene       51 CGGCKTTTGGTTTAGCCGTTTCAGTTTCAAACGTTTCATTTGTATCGGGA 100
                  **** *********************************************

AZM4_68973    100 ATTCTGATATAGCTTCGACTGATCCTATATTATTCCTGTGCTAGTACGTT 149
CA2 Gene      101 ATTCTGATATAGCTTCGACTGATCCTG---------------------- 127
                  **************************

AZM4_68973    150 TGATTTTTTTTTCTTCATTTCACAG 174
```

O. Alignment of AZM4_68973 and Exon 10 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810). The consensus translated sequence is also shown.

```
AZM4_68973   --C AAG AAG CCG GAG CTT TTC GGG CCT CTC AAG TCC GGC CAG GCC CCC AAG
gi:606810    --C AAG AAG CCG GAG CTT TTC GGG CCT CTC AAG TCC GGC CAG GCC CCC AAG
                  K   K   P   E   L   F   G   P   L   K   S   G   Q   A   P   K
```

P. AZM4_68973 sequence corresponding to Intron 10.

```
GTATGCGCTATTGCCTACTAGCCTATACTCCATTCTTATTCTTCTGAACCAAATGCATGCGCCCCGCGCGCGTGCTAATTGCTAAC
CCATGTGCTGCCATATATGCTAAGCTGGCGAGACTTGCATTTGCTTGGTAAATTATTGAGATGCCGCCGTCCCTATATAGGCTCAC
TTCCTAGTATATAGAACCTGGCGTGCCAGAATATTGCAAGTAACCAAGTACAGAGTTTATTGTTTTTCTTTATGGGTGTTCTGAGT
TGGCATCTATCCCATGCGCATGATTATTTCATGCATGCGTTCATGCTTTTAGCGGGTTCTACTAGTTTTGTTATCCATAAAAATTA
CCATATTTTAAAACTTCTTTTGAAAAAAAAAATTATATGTATCCTTGTGAAAGTCGACATTAGACCTAGTATATCGGCGTAGTCTA
CGCTACCGACATAACACGTATCGGCGCCATATAGATCAAGGAGCTCAGCCATGATATATATATATACTAATTGGATGACCTGTGGG
GATGGCATTGTCGCTGCATAGCTAACAACCGCGGGAACCGGCCTGATTTTTTGTGCTCCTTCTTTTTGCCTGACCTGACATGACAG
TGATTTTGCTATGCTGCATGCAG
```

Q. Alignment of AZM4_68973 and Exon 11 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810). This exon corresponds specifically to Repeat C of gi:606810, rather than Repeat B of gi:606814. Differences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68973   TAC ATG GTG TTC GCT TGC TCC GAC TCC CGT GTG TGC CCA TCG GTG ACC CTG GGC CTG
gi:606810 TAC ATG GTG TTC GCC TGC TCC GAC TCC CGT GTG TGC CCG TCG GTG ACC CTG GGC CTG
           Y   M   V   F   A   C   S   D   S   R   V   C   P   S   V   T   L   G   L

4_68973   CAG CCC GGC GAG GCC TTC ACC GTT CGC AAC ATA GCC GCC ATG GTC CCA GGC TAC GAC
gi:606810 CAG CCC GGC GAG GCC TTC ACC GTT CGC AAC ATC GCC GCC ATG GTC CCC GGC TAC GAC
           Q   P   G   E   A   F   T   V   R   N   I   A   A   M   V   P   G   Y   D

4_68973      AAG
gi:606810    AAG
              K
```

R. AZM4_68973 sequence corresponding to Intron 11.

```
GTATATATACACACTGACGATTGTGAACAACGCAATGGTCTCAATTTCTACTCACACGGCCGGCCGCGGCCTCTCGTTTTCGTGTC
GACTGCAG
```

S. Alignment of AZM4_68973 and Exon 12 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810). This exon corresponds specifically to Repeat C of gi:606810, rather than Repeat B of gi:606814. Differences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68973   ACC AAG TAC ACC GGC ATC GGG TCC GCC ATC GAG TAC GCT GTG TGC GCT CTC AAG GTG
gi:606810 ACC AAG TAC ACC GGC ATC GGG TCC GCC ATC GAG TAC GCT GTG TGC GCC CTC AAG GTG
           T   K   Y   T   G   I   G   S   A   I   E   Y   A   V   C   A   L   K   V

4_68973   GAG GTC CTC GTG GTC ATT GGC CAT AGC TGC TGC GGT GGC ATC AGG GCG CTC CTC TCC
gi:606810 GAG GTC CTC GTG GTC ATT GGC CAT AGC TGC TGC GGT GGC ATC AGG GCG CTC CTC TCA
           E   V   L   V   V   I   G   H   S   C   C   G   G   I   R   A   L   L   S

4_68973      CTC CAG GAC GGC GCA CCT GAC ACC TT
gi:606810    CTC CAG GAC GGC GCA CCT GAC ACC TT
              L   K   D   G   A   P   D   N  (F)
```

T. AZM4_68973 sequence corresponding to Intron 12.

```
GTAAGTCGCGACAGTAAAATATATACAAGTTTCATTTAGATATAAAAAACTATTTGCGCTTATTTATGTCATGCATGATTTTGATC
CTCTCTATACCATGTTGTGTGTTGGTTTGGTGTGGTGTACGTACGCAG
```

U. Alignment of AZM4_68973 and Exon 13 of the CA cDNA sequence (Burnell and Ludwig, 1997; gi:606810). This exon corresponds specifically to Repeat C of gi:606810, rather than Repeat B of gi:606814. Differences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68973   --C CAC TTC GTC GAG GAC TGG GTT AAG ATC GGC TTC ATT GCC AAG ATG AAG GTA AAG
gi:606810 --C CAC TTC GTC GAG GAC TGG GTT AAG ATC GGC TTC ATT GCC AAG ATG AAG GTA AAG
               H   F   V   E   D   W   V   K   I   G   F   I   A   K   N   K   V   K

AZM4_68973   AAA GAG CAC GCC TCG GTG CCG TTC GAT GAC CAG TGC TCC ATT CTC GAG AAG
gi:606810    AAA GAG CAC GCC TCG GTG CCG TTC GAT GAC CAG TGG TCC ATT CTC GAG AAG
              K   E   H   A   S   V   P   F   D   D   Q   C   S   I   L   E   K
```

V. AZM4_68973 sequence corresponding to Intron 13.

```
GTATGTTGTACATTCGTCGAGCAGTTACTGTTGCATGAATAGATTGGTTTTTGCTCACCAAAAGGACCTCTATTGTTTCTGCAG
```

W. Alignment of AZM4_68973 and the last exon of the CA cDNA sequence, Exon 14 (Burnell and Ludwig, 1997; gi:606810 and gi:606814). This exon corresponds specifically to Repeat C of these cDNA sequences, and the last 221 bp are the 3′-untranslated region. Differences are indicated with bolding. The consensus translated sequence is also shown.

```
4_68973   GAG GCC GTG AAC GTG TCC CTG GAG AAC CTC AAG ACC TAC CCC TTC GTC AAG GAA GGG
606810/4  GAG GCC GTG AAC GTG TCC CTG GAG AAC CTC AAG ACC TAC CCC TTC GTC AAG GAA GGG
           E   A   V   N   V   S   L   E   N   L   K   T   Y   P   F   V   K   E   G

4_68973   CTT GCA AAT GGG ACC CTC AAG CTG ATC GGC GCC CAC TAC GAC TTT GTC TCA GGA GAG
606810/4  CTT GCA AAT GGG ACC CTC AAG CTG ATC GGC GCC CAC TAC GAC TTT GTC TCA GGA GAG
           L   A   N   G   T   L   K   L   I   G   A   H   Y   D   F   V   S   G   E

4_68973   TTC CTC ACA TGG AAA AAG TGA AAA ACT AGG GCT ACG GCA ATT CTA CCG GCC CGC CGA
606810/4  TTC CTC ACA TGG AAA AAG TGA AAA ACT AGG GCT AAG GCA ATT CTA CCG GCC CGC CGA
           F   L   T   W   K   K   *

4_68973   CTC CTG CAT CAT CAT AAA TAT ATA TAC T-A TAC TAT ACT ACT ACG TAC CTA CCG ATA
606810/4  CTC CTG CAT CAT CAT AAA TAT ATA TAC TCT AAC TAT ACT ACT ACG TAC CTA CCG ATA

4_68973   TGC ACC CGA GCA ATG TGA ATG CGT CGA GTA CTA TAT ATC TGT TTT CTG CAT CTA CAT
606810/4  TGC ACC CGA GCA ATG TGA ATG CGT CGA GTA CT- --- ATC TGT TTT CTG CAT CTA CAT

4_68973   ATA TAT ACC GGA TCA ATC GCC CAA TGT GAA TGT AAT AAG CAA TAT CAT TTT CTA CCA
gi:606810 ATA TAT ACC GGA TCA ATC GCC CAA TGT GAA TGT AAT AAG CAA TAT CAT TTT CTA CCA

4_68973      CTT TTC ATT CCT AA
gi:606810    CTT TTC ATT CCT AA
```

X. The 3′-end of the AZM4_68973 sequence aligned with the 3′-end of the genomic DNA library clone.

```
AZM4_68973        CGCTGAGCTTTTTATGTACTATATCTTATATGATGAATAATAATATGACCGCCTTGTGAT
Glib_Clone1       CGCTGAGCTTTTTATGTACTATATCTTATATGATGAATAATAATATGACCGCCTTGTGAT
                  ************************************************************

AZM4_68973        CTAAAGACATCAGCTATATTTTTTCACAATATTATTACGAAGAGCTTCTTAGCTTTGTT
Glib_Clone1       CTAAAGACATCAGCTATATTTTTTCACAATATTATTGCGAAGAGCTTCTTAGCTTTGTT
                  ************************************ ***********************

AZM4_68973        AATTACCATTAGCGGATCTAGAAACGACCGAGGGGCAAAAGAATAGGACTTTCTTGGGAA
Glib_Clone1       AATTACCATTAGCGGATCTAGAAACGACCGAGGGGCAAAAGAATAGGACTTACTTGGGAA
                  *************************************************** ********

AZM4_68973        GCCAGTAAAGCAAGAGGTGCT----AAACACAGGGATAAAAGAACCCATATAAGCAACTA
Glib_Clone1       GCCAGTAAAGCAAGAGGTGCTTGCTAAACACAGGGATAAAAGAACCCATATAAACAACTA
                  *********************    **************************** ******

AZM4_68973        AGAAGATAACTAAAATAATATTCCTATGGATTACCTACCTAGGAAAAAGTCTTGAGATCT
Glib_Clone1       AGAAGATAACTAAAATAATATTCCTATGGATTACCTACCTAGGAAAAAGTCTTGAGATAT
                  ********************************************************** *

AZM4_68973        CTGGAGTTTCCAAATTAGACCTATAAGGTTAAAATTCATACTTACCAAATACTTATAGAT
Glib_Clone1       CTGGAGTTTCCAAATTAGACCTATAAGGTTAAAATTCATACTTACC--------------
                  **********************************************

AZM4_68973        CTAACAAACAATGCTCAATTCAAAGTGCTTAATTAAACAATATATAATTAATTATAGAAA
Glib_Clone1       -TAACAAACAATGCTCAATTCAAAGTGCTTAATCAAACAATATATCATTAATGATAGAAA
                   ******************************** *********** ****** *******

AZM4_68973        ATACCTAATAATAGCTCTTGATAACTAAATAATCAAATATTTCTTCACAATTTCAAACTA
Glib_Clone1       ATACCTAATAATAGCTCTTGATAACTAA---ATCAAATATTTCTTCACAATTTCACACCA
                  ****************************   ************************ ** *

AZM4_68973        CCTGGGCCGGTGTCCTCCATTGTAGATATGCCTATAATCATTGTGTAGATACTGATCAAG
Glib_Clone1       C-TGGGCCGGTGTCCTCCATTGTAGATATGCCTATAATCATTGTGTAGATACTGATCAAG
```

```
                 * **********************************************************

AZM4_68973       GGTCTCCTACCCTTATATTATATAAGCCAAGGAGAGGGTTACAAAATAT-----------
Glib_Clone1      GGTCTCCTACCCTTATATTATATAAGCCAAGGAGAGGGTTACAAAAGATATGATCAGCTA
                 ********************************************** **
```

3.7: Sequence alignment of AZM2_23203 and AY109272 (gi:21212748).

A. Sequence of AZM2_23203 68 bp upstream of the start of the AY109272 transcribed sequence.

```
GGGGGGGGGGGGGGAGAAGAAAAGAGTAGAAGGTGAGCAGCTGCCTTTTCCCCTCCGCCTCCGCCGTT
```

B. Alignment of AZM2_23203 and AY109272 upstream of the start codon. Differences in the sequences are indicated with bolding. The predicted translation, based on the AY109272 sequence is also shown.

```
2_23203   ATA AAA TGG GAG GAG GGA GGG GGG CAG CGA AGC CAT CCA TCC ATC CAT AGA TCC CTC
AY109272  GCA CGA GGG GAG GAG GGA GGG GGG CAG CGA AGC CAT CCA TCC ATC CAT AGA TCC CTC
           A   R   G   E   E   G   G   G   Q   R   S   H   P   S   I   H   R   S   L

2_23203   ACC TCA CTC ACT CGT GAA GAA CCA GAG AAC TCA CCC ACC CTC CTT CAT CTC CTA CAT
AY109272  ACC TCA CTC ACT CGT GAA GAA CCA GAG AAC CCA CCC ACC CTC CTT CAT CTC CTA CAT
           T   S   L   T   R   E   E   P   E   N   P   P   T   L   L   H   L   L   H

2_23203      CCA CAT CTA TCC AAG AAG
AY109272     CCA CAT CTA TCC AAG AAG
              P   H   L   S   K   K
```

C. Sequence corresponding to the first intron of AZM2_23203.

```
GTCCGTTCTCTCCTCCCCGCTCCCCTCAGCAATAATCTCCGCACGGCAACCTCGCTCCTCCGTTTTGGTTTTCCTCTGTGTGTGCG
TGCATTCTTTTCTCTTGTGCGTGCATGACGAGCTCTGACCGACGACGACGATGTTTAATTAACCTCTGGCTGTTGGCGACATTGCA
G
```

D. Alignment of the putative first exon of AZM2_23203 and AY109272. Differences in the sequences are indicated with bolding. The predicted translation, based on the AY109272 sequence is also shown.

```
2_23203   GCC GCC ATG GGC GAC GCC GTG GAG CAC CTC AAG AGC GGG TTC CAG AAG TTC AAG ACC
AY109272  GCC GCC ATG GGC GAC GCC GTG GAG CAC CTC AAG AGC GGG TTC CAG AAG TTC AAG ACC
           A   A   M   G   D   A   V   E   H   L   K   S   G   F   Q   K   F   K   T

2_23203      GAG GTG TAT GA
AY109272     GAG GTG TAT GA
              E   V   Y   D
```

E. Sequence corresponding to Intron 2 of AZM2_23203.

```
GTAAGTGCTTTGCTTGATTCCCCGATCGAGCAGTTTTCGCAGCCCCATGTCGTCGTTTGCTTATTGGTTTGTCTCTTGGGAATGAA
TTAATTCTGCGAGCTCCAACTGATCCCGTCGTTGCCCCCCTTTCGATCTTCTTCTTTTTCTTTCTCGCAG
```

F. Alignment of Exon 2 of AZM2_23203 and AY109272. The predicted translation, based on the AY109272 sequence is also shown.

```
AZM2_23203   C AAG AAG CCG GAG CTG TTC GAG CCT CTC AAG GCC GGT CAG GCC CCC AAG
AY109272     C AAG AAG CCG GAG CTG TTC GAG CCT CTC AAG GCC GGT CAG GCC CCC AAG
                K   K   P   E   L   F   E   P   L   K   A   G   Q   A   P   K
```

G. Sequence corresponding to Intron 3 of AZM2_23203.

```
GTACGTGACTAGTACTACTCGCCCTGTAGCTCTCTCTCTATCGTCACCGGCCCAGCATTGCCGCGTTCTCCCATCTTTCTGAACTG
CAAGGACGACCCTGCCGCACCTACGGCGCTGCTCCATCATCCACACACCCCAGCTGTGTCGCTGGGTGCGCAACTGCATTATTTGC
GCGTGTGACGAGGTCAACGCGTTGCGGTAGACAAACGGAAACAAGTAGTAGTCTGTTGGATTGGATATGATATGCGTACATAACTA
GCTTATACATATTTGGTTCGTACTTTATTTGTTATATTTTAGTTCAAAATAAACTAGCGGGTGATAAATATTCAAAAGTGCATGTA
GCGCCGGATATTTATACGCGTACGTGCATGAACCGTGGACACTAACGTGCATGCGCGCCGGCCGCCGAGGTGCATCCATCCCCATG
TGCAAATCCACCGGCCGGTAGTAGTCAAATATTCTTTGCACCAAAAGAAAAAAAAAGACAGAGGCAGTCTAGCCGTGTTGAATTTG
AATTACAATGGTAGTCTCACGTCCACACCTACCTAGCCGTGTTGAATTACGTACATTATTTGTTTTTGAACATATATATATGGTCA
CGTACGTCCATGTCCACATTTAATTTGAACGCCAGCAAACAGGTGGTGGTCCCGTCTGATTTTGGTTCGATGCATGTCTTAGCGTG
TGGTGGTCCAGCCATGATTGAGCGCAGATAGGCCCGAGCCCGGCCCCCACCATTCATTCAGCTCTCGACCGGTCCTGTGTCCGGTG
GCTGTGAAGCACTGGCCGGCACGTTGCTCTCTACTATACTCGAGAGTCGAGAGTAGTTGGGCGCATAATTGTTGGGCAGGGCCCAC
TGCCCCGTTGCCTTTTAGCATTGAGAGAACAGACACGCGGTTGCATCTGCCGTGCTGATATGGACACCCACGTCCCAGGTCCACAT
GACGAAATTCGTTATCGGGTTCATCTGTGTTCGTGGATATATGCAGCGCCTCACCATCTTTGCGTGCCAACTTCAACTTGGCGCGC
ACGGAAGCCAGCGTCGAATATCCTCGTCGCAAATAGTAGTAYGTCGGCTTGTTTGGTTTGCATCCACACATGTTGCACCTATTTTT
GTAATTAATGAAGATAAAATACAATACAAGTCATAGCCAGATAAAGTGCATGTTTGGTTGGCTGTTTAATTTGCCATATTTTATCG
CACTTTTGTGTGTAAACTTAGTTATTCAATTCGAAGGACTAAACTTACGGTAAAGTGTGATATAGTTATCCACCAACCAAACAGAC
TCCTCTAAAACCACGCACGGAACCGACCTTCCTGCGGCCTGGATTTTCTACGATCTGCCCCTGTATGCATATATATAAAATATACG
TAGTGAAAAGAGCGTCGGAACAAAGACAGGGGGCGGGTCAAAACTCAAAAGCCGCCGCCTTTTTGGTTCTGGGTGGCGTGCCGGCG
TGCAAATCCGCTGATCAGGGCCGATCTGCCTACCCGCTTATCGATAGGAGCCGCCGCGCGCGGGAACTGGACTGCATTTTTTTTCT
ATTATTTTTGGACTCCCTCTTCTTTTCGCCGACCTGAGACGGTGATTTCCCCTCTCTTCCRGTGCAG
```

H. Alignment of Exon 3 of AZM2_23203 and AY109272. The predicted translation, based on the AY109272 sequence is also shown.

```
2_23203   TAC ATG GTG TTC GCC TGC TCC GAC TCC CGC GTG TGC CCG TCG GTG ACC CTG GGC CTG
AY109272  TAC ATG GTG TTC GCC TGC TCC GAC TCC CGC GTG TGC CCG TCG GTG ACC CTG GGC CTG
           Y   M   V   F   A   C   S   D   S   R   V   C   P   S   V   T   L   G   L

2_23203   CAG CCC GGC GAG GCC TTC ACC GTG CGC AAC ATC GCC GCC ATG GTC CCG GCC TAC GAC
AY109272  CAG CCC GGC GAG GCC TTC ACC GTG CGC AAC ATC GCC GCC ATG GTC CCG GCC TAC GAC
           Q   P   G   E   A   F   T   V   R   N   I   A   A   M   V   P   A   Y   D

2_23203      AAG
AY109272     AAG
              K
```

I. Sequence corresponding to Intron 4 of AZM2_23203.

```
GTAACGCGCGCGCGCGCTCGTCGTACACACGCCACCCGGCGGCGCAATGGACGACGTTCTCACAACAACTCACTAGCAGCTACT
GTGTTGCCTTTCGTTACTTACCATCAATCAATAATTGCATGCAG
```

J. Alignment of Exon 4 of AZM2_23203 and AY109272. The predicted translation, based on the AY109272 sequence is also shown.

```
AZM2_23203 ACC AAG TAC ACC GGC ATC GGG TCC GCC ATC GAG TAC GCC GTG TGC GCC CTC AAG GTG
AY109272   ACC AAG TAC ACC GGC ATC GGG TCC GCC ATC GAG TAC GCC GTG TGC GCC CTC AAG GTG
            T   K   Y   T   G   I   G   S   A   I   E   Y   A   V   C   A   L   K   V

AZM2_23203 GAG GTC CTC GTG GTC ATC GGC CAC AGC TGC TGC GGT GGC ATC CGG GCG CTG CTC TCC
AY109272   GAG GTC CTC GTG GTC ATC GGC CAC AGC TGC TGC GGT GGC ATC CGG GCG CTG CTC TCC
            E   V   L   V   V   I   G   H   S   C   C   G   G   I   R   A   L   L   S

AZM2_23203   CTC CAG GAC GGT GCA CCC GAC AAC TT
AY109272     CTC CAG GAC GGT GCA CCC GAC AAC TT
```

```
              L   Q   D   G   A   P   D   N  (F)
```

K. Sequence corresponding to Intron 5 of AZM2_23203.

```
GTAAGTATCTCGTCGTTCACAACCATACATTTCTGATCATCTATCCGATCGAGGCGTACGTGTGCCCATGCATGCATGGTGTGTTT
ATGTCATGGTGTCGTCATGGCATGCAG
```

L. Alignment of Exon 5 of AZM2_23203 and AY109272. The predicted translation, based on the AY109272 sequence is also shown.

```
2_23203   --C CAC TTC GTC GAG AAC TGG GTC AAG ATC GGC TTC CCT GCC AAG GTC AAG GTG AAG
AY109272  --C CAC TTC GTC GAG AAC TGG GTC AAG ATC GGC TTC CCT GCC AAG GTC AAG GTG AAG
               H   F   V   E   N   W   V   K   I   G   F   P   A   K   V   K   V   K

2_23203      AAA GAG CAC GCC TCC GTG CCC TTC GAT GAC CAG TGC TCC ATC CTG GAG AAG
AY109272     AAA GAG CAC GCC TCC GTG CCC TTC GAT GAC CAG TGC TCC ATC CTG GAG AAG
              K   E   H   A   S   V   P   F   D   D   Q   C   S   I   L   E   K
```

M. Sequence corresponding to Intron 6 of AZM2_23203.

```
GTACGTACATACACCGCCGTACACGTGCGATCGAGTAGTTTATTGGTACGTTTCGTGATTTGCTCACCAGAGAGGACGATCGACCT
TGTGTTTCTGCTGCAG
```

N. Alignment of Exon 6, the last exon of AZM2_23203 and AY109272. The predicted translation, based on the AY109272 sequence is also shown until the stop codon.

```
2_23203   GAG GCC GTG AAC GTG TCC CTG GAG AAC CTC AAG ACC TAC CCC TTC GTC CAG GAA GGG
AY109272  GAG GCC GTG AAC GTG TCC CTG GAG AAC CTC AAG ACC TAC CCC TTC GTC CAG GAA GGG
           E   A   V   N   V   S   L   E   N   L   K   T   Y   P   F   V   Q   E   G

2_23203   CTG GCC AAG GGG ACC CTC AAG CTG GTC GGC GCC CAC TAC GAC TTC GTG TCC GGG AAC
AY109272  CTG GCC AAG GGG ACC CTC AAG CTG GTC GGC GCC CAC TAC GAC TTC GTG TCC GGG AAC
           L   A   K   G   T   L   K   L   V   G   A   H   Y   D   F   V   S   G   N

2_23203   TTC CTC GTA TGG GAA ACT TAG CCA CAC CCG CCG CTC AGC GCG CGG CTT CGT CGT CAC
AY109272  TTC CTC GTA TGG GAA ACT TAG CCA CAC CCG CCG CTC AGC GCG CGG CTT CGT CGT CAC
           F   L   V   W   E   T   *

2_23203   AAC TCC TAT ATC TTC ATA AAT ATA TAT ATC CAC TAC CTA TAT ATA CCT ACC GAT ATG
AY109272  AAC TCC TAT ATC TTC ATA AAT ATA TAT ATC CAC TAC CTA TAT ATA CCT ACC GAT ATG

2_23203   TGA ATG CGA GGA GTA CTA TTA CCT GCA TGT TTT CAG CAG TAT GCA GCT ACG TAC ATA
AY109272  TGA ATG CGA GGA GTA CTA TTA CCT GCA TGT TTT CAG CAG TAT GCA GCT ACG TAC ATA

2_23203   TAT ATG TAC CGG ATC GCC TAA CGA CGT GAA TGT AAT AAG CAA TAA TAT TTT CTA CCG
AY109272  TAT ATG TAC CGG ATC GCC TAA CGA CGT GAA TGT AAT AAG CAA TAA TAT TTT CTA CCG

2_23203      CTT TTT TTT AAA AAA AAA AAA AAA AAA
AY109272     CTT TTT TTT
```

O. Sequence corresponding to the 3′-end of AZM2_23203.

```
ATTTCTTTCTAACCTCGAGCTTCTATATGTAGCAACAACATTTTCTACGTACATATATATGTGCCACCCGTATTCCGACTGATATA
CCTATCACTAGTACAAAAGAGATCTAAGCTACCAGTACGCACAAATTTTCACTGACGGTTTTAATTGGTGGTTCCTTAATCTAACT
ATCCGTAAACTTTATGTTATAAATACACTTTCTTCCTCCAACATGAATAGTAGGTCACAGCCAACGGAAGAGGCACACCAATATTC
GAGTAACCGGGCCGGATTGGGCCGTATAGTTGGGCTGCGTCTTGTAAAGGATTACCTTTTGCGTCGGCCTGCCTATAAAAAGAAGC
GGATGACTTCACCTCCGGGCATGATGAATGAGAATAACAAAGTTCCAACCCCTACATTCTTCTTCTACCTCTGCCATCCTCTCTGT
ATTTTTCCCAGCTCCTATCTTCTCTTCGTCGTAGTGCTGCCCTTCAGCGAGGTGCTGGACTGGAGCACCTTCTCGGTGGTGCTCGC
CGAGAGAAGGGCATACCGGCGGACCTCAAGAACTTCCTGTGGCACGCCGCACGCCACGCCCATCAAGTACGACTTCTTCCACATGA
```

```
TCCTGCACTCCATCTGGTTGAGTAGAGCGAACCAGGTCGAGCTCGACGGCTGAACTGTGCGGTCAAGTCTCTGGAACAGATTAGTC
TGTAACTGAGCTCGACGGCTGATCTTACGGCAGCTTGCCCTACTGTGCTGTCAGTTAGGCTGTTCAGATGCGCCCGCACATGTAAA
CAATTCTATATATGCAACCTTTTTTTGCATCATACTCTAATTCTTAGAATAAACGAAAATTTTGTTTCCATCCAACTAACTCATTG
AATGTTGCAACTCTAGCTCGCTGATCCAACACTTAAAAAAAGCAGCAGCGCAGCTTACTTTGAATGTTATCCGTTTGAAGCAGAAT
CAACTCCACGAGCATTTTATCCTTTAATTAATTCATCTCTTCCTGTTGTTGGGGGGGTGGGGGGGGGGGTCCCCTACAGCCTTTAC
TGTTAGAAAAAACATTTCACTGACAAAGAAGTGATATAAAATCATAAACCAGTGCCAAAAACAGCGTGTGCAACAGCTAATTAACG
CGCGGTTACATTTATTTCCAGAATCTACAATTAGTACTTGATTCCAACAACGCCAGAGCAGGACGAGTCGTGTCAGACCATGCTGA
AGAGATTAGCCGTGGAGATGACCGCGTTCAGAGCGAGCACCAGGAACGCGACGAAGGACAGCCA
```

Appendix - Chapter 4

4.1: Analysis of transcription factor binding sites in the CA2 gene sequence

A. ClustalW alignment of the AZM4_68974 assembly with the CA2 gene sequence in the region 2.5 kb upstream of the start codon (ATG). Putative TATA boxes are shown with bolding, potential transcription start sites are boxed, CCAAT-boxes (and inverted) are shown in red, possible transcription factor-binding sites are numbered, and details are provided in Table 4.10. Regions homologous (50% in 20 base pair window and 70% in 15 base pair window) with the rice CA genomic sequence are underlined with red.

```
2 Sequences Aligned           Alignment Score = 6486
Gaps Inserted = 4             Conserved Identities = 953
Pairwise Alignment Mode: Fast
Pairwise Alignment Parameters:
ktup = 2   Gap Penalty = 5   Top Diagonals = 4   Window Size = 4
Multiple Alignment Parameters:
Open Gap Penalty = 20.0   Extend Gap Penalty = 5.0
Delay Divergent = 40%     Transitions: Weighted


AZM4_68974         1 CTACTGCTACCTTAAAATATGCATGCATGCACGACAAAGGGTAAACGCGC   50

AZM4_68974        51 ATGACGCCACCAGTTTAATTTGTAGCTTTTGCATTTTCAATTCCCTTACC  100

AZM4_68974       101 TTTGTGTTGTCGCCGTCACGTACGCGCTGTCTAGCTAGCGATCTCATGTG  150

AZM4_68974       151 CTTATGATGTCGCCAAATGAGCACATCCTGGCAGCATGCCAAGGCTCAAG  200

AZM4_68974       201 TGTGTCAACCTTGCACGTACATGTACGTCGCCGGTTYCGTTCCGTCCACC  250

AZM4_68974       251 GGGCAGCACAATGAGCTTCAGCTAGCGCTGGGCCGTGCAAACCCTAACCG  300

                                           3c*             2b*
AZM4_68974       301 CCGCCCACCGCCATGGATCGGGCGGCATGCATAGAGTTCGTGGCGAGCGG  350

AZM4_68974       351 CGGAGCGCTAGGTAGAGCTTAGACCTCATCAGATTTGTTCTACTCTACTG  400

AZM4_68974       401 AATCCTGGTTGCATYTCATGGATGCATGTTTYTTCTTCTGCTTATTATTG  450

AZM4_68974       451 TCTGTGCTTCAAGCTCTGTCAATGCTATATATATATATGCTCCGCCACTG  500

AZM4_68974       501 CTCGTCGCTAATTCGATCGGCAGCGCGGGCATCGGAGCAGCTAGCCACGC  550

AZM4_68974       551 ATGCATCAGCCAAGCTTGGATGGAATGGAGCAGCTGGGACGGTTCCAGAT  600

AZM4_68974       601 GATTAAATATTATACTCCGTACAAATTATACAGCATCCCAAATGGATTCT  650

AZM4_68974       651 TTGTAACCGAACAAGTCAGGATGAATACGGCGGGCATCGGAGAAAGGTGA  700

AZM4_68974       701 GTTCATCCACTTCTCTCTCACAACTCTGTCAAGAGTGGTATATATATGTT  750

AZM4_68974       751 TAAACTATATTGCATATTCCAGTGAGACATGTCAGCTCTGGGGCGCTGCA  800

AZM4_68974       801 GCCTGGATCCAGAAGCTTCACAATTCTATACCACCCCTCACTTGTTCTGT  850

AZM4_68974       851 TGTTTGAAAACCAGGTTTTCGAACGACATTTCAAGCTCAATTAGCCCAGT  900

AZM4_68974       901 ACCCAGTTACCCCACTGGGGACACTTTGTACATGAGCAGCCAGCCAATC  950

AZM4_68974       951 TGCGCATAGGCTAACGCCTAACGGCCTGGCCCCGCCCATGTCAGCAGGCC 1000



                                 4b
AZM4_68974      1001 TCCGAGGCTTTTGGTTGCCCAACCAGCCCATGGGCTGAATTCATAACAGT 1050
CA2 gene           1                                     GAATTCATAACAGT   14
                                                         **************
```

```
AZM4_68974       1051 GTTGGCACACAGTTTCCTCTTCACTCGGAAGCTTATTATTATCGATCCTG 1100
CA2 gene           15 GTTGGCACACAGTTTCCTCTTCACTCGGAAGCTTATTATTATCGATCCTG   64
                      **************************************************

                                                   14
AZM4_68974       1101 AACCAGAGACTAGCAGAGCTAGCATTTCGACGACGCGTCTCAACTCTCAA 1150
CA2 gene           65 AACCAGAGACTAGCAGAGCTAGCATTTCGACGACGCGTCTCAACTCTCAA  114
                      **************************************************

                             11/2b*   2b  3c 5    3b     4a      5
AZM4_68974       1151 CCTCCAAGTCCACCTCGTGTACGTGCTGCCTTGCCAGTTGCCACTGGGCA 1200
CA2 gene          115 CCTCCAAGTCCACCTCGTGTACGTGCTGCCTTGCCAGTTGCCACTGGGCA  164
                      **************************************************

                                     2a    2a/4b           11/3c*/3a
AZM4_68974       1201 CTGCTGGCCCAGTGACCAACCATGCGTTAGATCTGACAGCACCACCGAAC 1250
CA2 gene          165 CTGCTGGCCCAGTGACCAACCATGCGTTAGATCTGACAGGACCACCGAAC  214
                      *************************************** **********

                                     8/4b
                             1   3a  /2a   2b*/14
AZM4_68974       1251 CATCCTCCCCGGTGATCAACAAACGACGGCAGCCACATCTTGCACCCAAC 1300
CA2 gene          215 CATCCTCCCCGGTGATCAACAAACGACGGGAGCCACATCTTGCACCCAAC  264
                      ***************************** ********************

                      3b/5                    9      8/4b
AZM4_68974       1301 GTGATGATGAATGATGCCTAGAACTTTTGACAACAAAACGCAGCACAGGT 1350
CA2 gene          265 GTGATGATGAATGATGCCTAGAACTTTTGACAACAAAACGCAGCACAGGT  314
                      **************************************************

                          5/6        8
AZM4_68974       1351 AGCAGGTTTAATTCAACAAGACTTTCTACTATATAGAGCCACACCATAGA 1400
CA2 gene          315 AGCAGGTTTAATTCAACAAGACTTTCTACTATATAGAGCCACACCATAGA  364
                      **************************************************

                          4b                9               3a   4a/11
AZM4_68974       1401 GATAACTAATCTGTGCGCAAAGCCAAAGTGCTGAC----GGCAACTGTGG 1446
CA2 gene          365 GATAACTAATCTGTGCGCAAAGCCAAAGTGCTGACTGACGGCAACTGTGG  414
                      ***********************************    ***********

                             9        4b   6/4b   15           9  4b
AZM4_68974       1447 TGCAGCCTTTTCATCTCCGTTTTTAAGTTTTTTGCCCCTCCTTTTGTTTT 1496
CA2 gene          415 TGCAGCCTTTTCATCTCCGTTTTTAAGTTTTTTGCCCCTCCTTTTGTTTT  464
                      **************************************************

                       4b             9   6 5/3a   1    3a
AZM4_68974       1497 CTGTTTTTCTGGGAACTCTTTAAACCGCCGTGGCGCCGTGTAAACTTTGC 1546
CA2 gene          465 CTGTTTTTCTGGGAACTCTTTAAACCGCCGTGGCGCCGTGTAAACTTTGC  514
                      **************************************************

                             9
AZM4_68974       1547 TGTAGCCTTTTCGCGTGCAATGGCAGAGCGCCCTGTTCTTTTCCTGCTAA 1596
CA2 gene          515 TGTAGCCTTTTCGCGTGCAATGGCAGAGCGCCCTGTTCTTTTCCTGCTAA  564
                      **************************************************

                      9            9       8                 1/3a
AZM4_68974       1597 AGAA--AAAAAAAAAGGAGCACCTGATCGCTGGCAGGCCCACGGCCCACC 1644
CA2 gene          565 AGGAGAAAAAAAAAAAGGAGCACCTGATCGCTGGCAGGCCCACGGCCCACC  614
                      ** *  ********************************************

                       4a                3a             7
AZM4_68974       1645 CAACTGTGTCTGTAACGCTCGGCGTCCCTGCATTGCATGCCAAGTGCCAA 1694
CA2 gene          615 CAACTGTGTCTGTAACGCTCGGCGTCCCTGCATTGCATGCCAAGTGCCAA  664
                      **************************************************

                                                  3a          3a/1
AZM4_68974       1695 CCACCAGTCCATAGCAGGGTCAGGGAGACCGCAGATGAGGCCGGGGCAAC 1744
CA2 gene          665 CCACCAGTCCATAGCAGGGTCAGGGAGACCGCAGATGAGGCCGGGCCAAC  714
                      ********************************************* ****

                         3a/1      9    5       5           9/15
AZM4_68974       1745 GGTGATGCCGCAAAGAGGATTCAGAATCCTTTTTCTTTTCTTTTCTTTTA 1794
CA2 gene          715 GGTGATCCCCCAAAGAGGATTCAGAATCCTTTTTCTTTTCTTTTCTTTTA  764
                      ****** ** ****************************************
```

```
                    3c*/3a                    2b             3c/5   3b
AZM4_68974      1795 CCACCGGGCTGGCATCACAGATTACACGCGCAGTAGAGTAAGCACGTCTC 1844
CA2 gene         765 CCACCGGGCTGGCATCACAGATTACACGCGCAGTAGAGTAAGCACGTCTC  814
                     **************************************************

                                     8             3b
AZM4_68974      1845 TCTCGTAGCCAAGAACAACAGTCTA-CACAGCTCGCTTTCTCCGCCCTTG 1893
CA2 gene         815 TCTCGTAGCCAAGAAAAACAGTCTAGCACAGCTCGGATTCTCC-CCCTTG  863
                     *************** ********* *********  ****** ******

                                 3a                         3b
AZM4_68974      1894 TCTGGGCGTTACGGCAGGCAAGCCCCCTCGTTTTCTTCTGCTCGCGTTCT 1943
CA2 gene         864 TCTGGGCGTTACGGCAGGCAAGCCCCCTTGTTTTCTTCTGCTGGCGTTCT  913
                     **************************** ************* *******

                                            11/3c*/3a
AZM4_68974      1944 CCTTCCATGTCCACATCTCCTGTGCCACCGCACGCAAGGTGCCAACGCTC 1993
CA2 gene         914 CCTTCCATGTCCACATCTCCTGTGCCACCGCACGCAAGGTGCCAACGCTG  963
                     *************************************************

                         1/3a                            11
AZM4_68974      1994 CCTCGCCGCAGTAGCATCGCGTCCACACAAACTGCACCTCCACTAGATAC 2043
CA2 gene         964 CCTCGCCGCAGTAGCATCGCGTCCACACAAACTGCACCTCCACTAGATAC 1013
                     **************************************************


                     3a/11
AZM4_68974      2044 GGCGGTGATCCGGCGAGAGAGCGCGACACGCACAGGCCAGCTAGCGTTTC 2093
CA2 gene        1014 GGCGGTGATCCGGCGAGAGAGCCCGATACGCACAGAACAGCTAGCGTTTC 1063
                     ********************** *** ********  *************

                           1/3a                               1/3a
AZM4_68974      2094 TCC--GACGCCGCGCGTTTCATCATTTCCCGCTTCCCCTGCCCCCGGCCG 2141
CA2 gene        1064 CTCACGACGCCGCGCGTTTCATCATTTCCCGCTTCCACTGCCTCCGGCNG 1113
                       *  ******************************* ***** ***** *

                                       2b                         7
AZM4_68974      2142 CG--CGCGCGCGCCCGTGTGGTCCAGACCAGGACGCGCGCGGATGTGCAT 2189
CA2 gene        1114 NGNGCGCGCGCGCCCGTGTGGTCCAGAACCAGGAGCCCG-GGATGTGCAT 1162
                      *  *********************** *  *  ** ** **********


                       3a/3c   14/3a/12b    3a   16c/1/3a/3c    3b
AZM4_68974      2190 CCGGCGCGCGCCCGTCGGCCACACGGTGCCGCCGCGCGTTATCCCGAGCC 2239
CA2 gene        1163 CCGGCGCGC-CCCGTCGGCAACACGGTACCGCCGCGCGTTATCCCGAACC 1211
                     ********* ********* ******* ******************* **


AZM4_68974      2240 CTGTCCTGTCCTGTCCTGTTCCATCTCGCGCGCGAGGGGGGGAGGGGAGG 2289
CA2 gene        1212 CTGTCCTGTCCTGTCCTGTTCCATCTCGCGCGCGAGGGGGGGAGGGGAGG 1261
                     **************************************************


                                                  3a           3c*/3a
AZM4_68974      2290 GCAGCGAGTGGCGCGCTGGCGGATGAGGCGCCGAGTGGCCCGCATCCACC 2339
CA2 gene        1262 TCAGCGAGTGGCACGCTGGCGGATGAGGCGCCGAGGTGCCCTCATCTACC 1311
                      *********** **********************  **** **** ***


                                          16c                 16c
                               1/3b  14/3c/3a/1    3b        3a/1/12b
AZM4_68974      2340 GGCGCAGGCGAGCCGCACGACGCCGCCGCGCTCGCGGACCGCCGCCGCCA 2389
CA2 gene        1312 GTCGCAGGCGATCCGCACGAAGCCGCCGCGCTCGCG-ACCGCCGCCGCCA 1360
                     * ********* ******** *************** *************


                     2b/4c     11/1                12c            1
AZM4_68974      2390 CACATGCGCACCCCCGGCCCGCGGGGCTGTAACGGCCTTGTCGCCACGCG 2439
CA2 gene        1361 CACATGCGCACCCCCGGCC-GCGGGGCTGTAACGGCCTTGTCGCCACGCG 1409
                     ******************* ******************************

                            2b
AZM4_68974      2440 TGCGCCCCGTGTGTATAAGGAGGCAGCGCGTACAGGGGGCACGATAAGC- 2488
CA2 gene        1410 TGCGCCCCGTGTGTATAAGGAGGCAGCGCGTACAGGGGGCACGATAAGCC 1459
                     *************************************************
```

```
AZM4_68974       2489 -------------------------------------------------- 2489
CA2 gene         1460 TTGTCACNANGCGTGCNCCCCGTTTGAATAAGGAGGCAAGNGCGTACAGG 1509

                                                   Start
AZM4_68974       2489 -------------GGCACTCGCACGATCAATGTACACATTGCCCGTCCGC 2525
CA2 gene         1510 GGGCACGATAAGCGGGACTCGCACGATCAATGTACACATTGCCCGTCCGT 1559
                                   ** *********************************
```

B. ClustalW alignment of AZM4_68974 with the CA2 gene sequence in the intron region including Exon 2 of the CA2 (gi:606810) sequence. Putative TATA boxes are shown with bolding, potential transcription start sites are boxed, CCAAT-boxes (and inverted) are shown in red, possible transcription factor-binding sites are numbered, and details are provided in Table 4.10.

```
CA2 gene            1 GTACGTACGTGCGTACGGTGTACGACAATTAATGCATGCATGGCTCACTG   50
AZM4_68974          1 GTACGTACGTGCGTACGGAGTACGACAATTAATGCATGCATGGCTCACTG   50
                      ****************** *******************************

CA2 gene           51 CACAGCGAGGACATCATGCACTGTACAGCATGCATGTCCTCGTTTCTATA  100
AZM4_68974         51 CACAGCGAGCACATCATGCACTGTACAGCATGCATGTCCTCGTTTCTATA  100
                      ********* ****************************************

CA2 gene          101 TATTCTATCCACGTACGCCTTCGCTTCATCCAGCTCTAATAACCAAGTAC  150
AZM4_68974        101 TATTCTATCCACGTACGCCTTCGCTTCATCCAGCTCTAATAACCAAGTAC  150
                      **************************************************

CA2 gene          151 TGACTAGTCGGCACTACTGACGACCTTGCTGTTTTGGGCACGAAATGCAC  200
AZM4_68974        151 TGACTAGTCGGCACTACTGACGACCTTGCTGTTTTGGGCACGAAATGCAC  200
                      **************************************************

                             4b   9                              2b*
CA2 gene          201 TCATGCAACCAAAAGTCCTTGCTCCCATTATCTAGGAGGACACGACCAAG  250
AZM4_68974        201 TCATGCAACCAAAAGTCCTTGCTCCCATTATCTAGGAGGACACGACCAAG  250
                      **************************************************

CA2 gene          251 CATGCAGGATATTCCTGGAAACCCCAATCAAAATCCGACTGGTAATATAA  300
AZM4_68974        251 CATGCAGGATATTCT--GGAAACCCAATCAAAATCCGACTG-TAATATAA  297
                      **************   * ** ******************* ********

                                                                    6
CA2 gene          301 GTATAATGCACATCCCCAAGAAGACGCCACTCTTAGCTTCATCTTTAAAT  350
AZM4_68974        298 GTATAATGCACATCCC-AAGAAGACGCCACTCTTAGCTTCATCTTAAATT  346
                      **************** **************************** ** *

                                                         6        15
CA2 gene          351 TCCCCCTTAAATTTAGCTATATATCAATTTATTTAACAATCAAAAACAGG  400
AZM4_68974        347 ---CCCTTAAATTTAGCTATATATCA-TTTATATAACATC--AAAACAGG  390
                         *********************** ***** *****    ********
                                 15
CA2 gene          401 CGTACTATCAAAAATTTTTTTTATGGTACGTAAATATGGCACACTACACT  450
AZM4_68974        391 CGTACTATCAAAAATATATTTCATGGTACGTAAATATGGCACACTACACT  440
                      *************** * *** ****************************

                                15/4b               10
CA2 gene          451 GTACATAAATTTTTGTTTGAATTAATTCTTTCTGTCAAATCTATAATCAA  500
AZM4_68974        441 GTACATAAATTTTTGTTTGAATTAATTCTTTCTGTCAAATCTATAATCAA  490
                      **************************************************

                                           4b/5
CA2 gene          501 ATTCAAGGCAGTTTGATATATACGTTAGATCTATATATAATGCATCTATT  550
AZM4_68974        491 ATTCAAGGCAGTTTGATATATACGTTAGATCTATATATAATGCATCTATT  540
                      **************************************************
```

```
CA2 gene         551 TCTGGCGGAGGGAATAGCTAGTGATGATGATAGTAATTTAGATCATTTTC  600
AZM4_68974       541 TCTGACGGAGGGAATAGCTAGTGATGATGATAGTAATTTAGATCATTTTC  590
                     **** *********************************************

                                       3a
CA2 gene         601 CCATCAGCTAGTAGCTACCGACGACATACGCATGTCAGCCATCTCCAATA  650
AZM4_68974       591 CCATCAGCTAGTAGCTACCGACGACATACGCATGTCAGCCATCTCCAATA  640
                     **************************************************

                                               15/9   16b
CA2 gene         651 GAATATTCCCGAAGGGAGGTGTTTCCAAAAAGATGACGGGCAAACGATAG  700
AZM4_68974       641 GAATATTCCCGAAGGGAGGTGTTTCCAAAAAGATGACGGCCAA-CGATAG  689
                     *************************************** *** ******

                                         5             9           4b
CA2 gene         701 TGCTAGTTTGAAGGCAACTACTACGTATATATCCTTTCAGTATAACAGAA  750
AZM4_68974       690 TGCTAGTTTGAAGGCAACTACTACGTATATATCCTTTCAGTATAACAGAA  739
                     **************************************************

                                   15/9             16b/9     16b/3b/5
CA2 gene         751 TTCCACCCAGAAAAAAAAAAGTCTCGAGTTGAATGAAAGAGGAGTAGTGAC  800
AZM4_68974       740 TTCCACCCAGAAAAAAAAAAGTCTCGAGTTGAATGAAAGAGGAGTAGTGAC  789
                     **************************************************

                        14               9                15     9
CA2 gene         801 GTCGAGCGCGCGTGAAATAAAGTATATGGCTGGCTTTTTCCTAAAGCGAT  850
AZM4_68974       790 GTCGAGCGCGCGTGAAATAAAGTATATGGCTGGCTTTTTCCTAAAGCGAT  839
                     **************************************************

                                                    2b/4c
CA2 gene         851 AAGACCAGTTTATGCAGTGGGGTCATGGACATGTGTAGTGATAGCTAATA  900
AZM4_68974       840 AAGACCAGTTTATGCAGTGGGGTCATGGACATGTGTAGTGATAGCTAATA  889
                     **************************************************

                                        16b/4         16b
CA2 gene         901 ATCGTCCGCGTCTTTTGGCTTTTGAGTTCCGTTTGATCCATGACGCATAT  950
AZM4_68974       890 ATCGTCCGCGTCTTTTGGCTTTTGAGTTCCGTTTGATCCATGACGCATAT  939
                     **************************************************

                                4a      4b/3a/14          9/3a/1/16c
CA2 gene         951 ATATCCAGGCAGTTGAATAACCGACGACCATCAAATAAAAGGCCGCCACT 1000
AZM4_68974       940 ATATCCAGGCAGTTGAATAACCGACGACCATCAAATAAAAGGC-GCCACT  988
                     ******************************************* ******


                                     5/3b/4   6
CA2 gene        1001 ACTAGTGGCCATCGACGTCAGTTTAACCTTTCTATGTATGCATGTGTAAC 1050
AZM4_68974       989 ACTAGTGGCCATCGACGTCAGTTTAACCTTTCTATGTATGCATGTGTAAC 1038
                     **************************************************

                                           4b           4b     3a
CA2 gene        1051 TTCCCATGATTTCCTTGGCTGCGTTATTTTGCTTTGTTTCACCGTCGGAC 1100
AZM4_68974      1039 TTCCCATGATTTCCTGCGTCGCGTTATTTTGCTTTGTTTCACCGTCGGAC 1088
                     ***************  *  ******************************

CA2 gene        1101 GACGAAGTCTTTTAGATAGCAATAAGGAACTATATCTAAGTCCTAGTTTG 1150
AZM4_68974      1089 GACGAAGTCTTTTAGATAGCAATAAGGAACTATATCTAAGTGCTAGTTTG 1138
                     ***************************************** ********

                                     1                            4b
CA2 gene        1151 GGAACCTCGTTTTCCCACGAGATTTTCATTTTCCTAAGGTAAATTAGTTC 1200
AZM4_68974      1139 GGAACCTCGTTTTCCCACGAGATTTTCATTTTCCTAAGGTAAATTAGTTC 1188
                     **************************************************

                                                         9
CA2 gene        1201 CGGCTTTTTTGAAAATAAGAATCTTTTGAAAAAGATGGTAATTATCAAAC 1250
AZM4_68974      1189 ATTTTTTTTTGAAAATAAGAATCTTTTGAAAAAGATG-TAATTATCAAAC 1237
                         ********************************* ************

                              4b        15                   9
CA2 gene        1251 TAGTCCTAACAGAGAGATTTTTGAGGGGGGAGAAAAAAAAGGAAGTTCTT 1300
AZM4_68974      1238 TAGTCCTAACAGAGAGATTTTTGAGGGGGGAGAAAAAAAAGGAAGTTCTT 1287
                     **************************************************

                                           4b
CA2 gene        1301 CTGCATTCTTTTTTGGAGGAACAAAAAATTTGCCTCTGCATACTGAATCA 1350
```

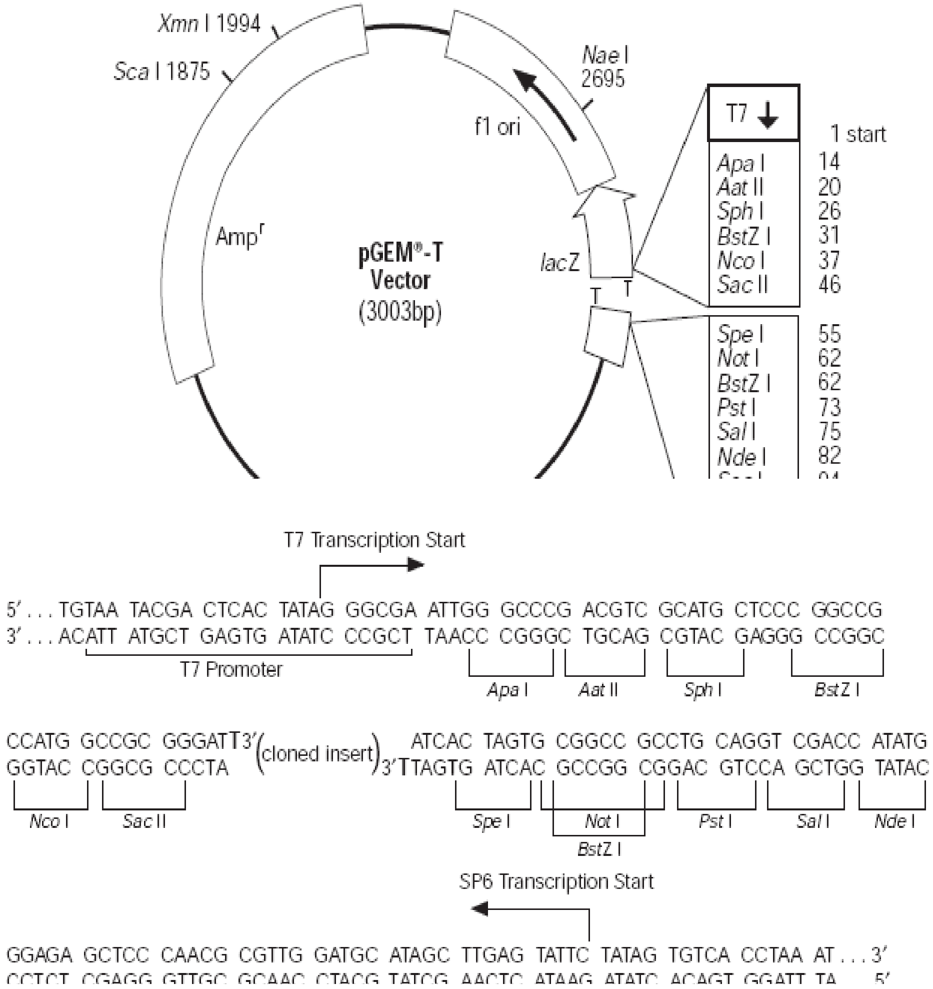
```
AZM4_68974   1288 CTGCATTCTTTTTTGGAGGAACAAAAAATTTGCCTCTGCATACTGAATCA 1337
                  **************************************************

                                      2b*
CA2 gene     1351 GAGGGGATGGCTTTATTTCGTGTTGGCTGGTTGATTGATGATTGGATGA 1400
AZM4_68974   1338 GAGGGGATGGCTTTATTTCGTGTTGGCTGGTTGATTGATGATTGGATGA 1387
                  **************************************************

                          4b                                 2a
CA2 gene     1401 GCTCCAGTAAGTTTGGAAGAGAACAGGGCACGGTCCCGACGGTTGGTACG 1450
AZM4_68974   1388 GCTCCAGTAAGTTTGGAAGAGAACAGGGCACGGTCCCGACGGTTGGTACG 1437
                  **************************************************

                            9          6
CA2 gene     1451 GGTGAAGAAAGGGAGTGATTTAATTTATCGCCC-AACCACAACCACCCAT 1499
AZM4_68974   1438 GGTGAAGAAAGGGAGTGATTTAATTTATCGCCCCAACCACAACCACCCAT 1487
                  ********************************* ****************

                                                          9/3a/1
CA2 gene     1500 CGATCTATAGTTGCAGAAGAACTCGCTAATGCTGTCCACAAAAGCCGCAC 1549
AZM4_68974   1488 CGATCTATAGTTGCAGAAGAACTCGCTAATCCTGTCCACAAAAGCCGCAC 1537
                  ****************************** *******************

CA2 gene     1550 TCACGCACTCATCCGCCACTGATTTTATTTCCCCCCCCCCCCCT----GT 1595
AZM4_68974   1538 TCACGCACTCATCCGCCACTGATTTTATTTCCCCCCCCCCCCCCTGTGGG 1587
                  *******************************************     *

                                   1                  15
CA2 gene     1596 GGCGCGCGGTTGCTGCGTGGTGGTACTACTACCTGTTTTTGCTCACTGAC 1645
AZM4_68974   1588 CGCGCGCGCGTGCTGCGTGGTGGTACTACTACCTGTTTGT-CTCACTGAC 1636
                   *******  **************************** * *********

                                             Exon2──▶   16a
CA2 gene     1646 ACAGTTGCGGGT-TCATCATGTTGCT>AGTAAACGGGACGGCGGGCAGCTG 1694
AZM4_68974   1637 ACAGTTGCGCGCGTCATCATGTTGCT>AGTAAACGGGACGGCGGGCAGCTG 1686
                  ********* *  *************************************

                      16d                         9/10
CA2 gene     1695 AGGAGTCAAACGAGAGAGATCGAGAGAGAAAGAAAGGGAGGGCATCCACC 1744
AZM4_68974   1687 AGGAGTCAAACGAGAGAGATCGAGAGA--AAGAAAGGAGGGCATCCACC 1734
                  ***************************  *********************

                    3a
CA2 gene     1745 AGCCGGCGGGCATAAGAGGGGAGGAGAGAGAGGCCAGAGAAGAGGAGGAG 1794
AZM4_68974   1735 AGCCGCCGGCGATAAGAGGGGAGGAGAGAGAGGCCAGAGAAGAGGAGGAG 1784
                  ***** ***  ***************************************

                                                          15/9
CA2 gene     1795 AAGAAGAAGAAGATGAGCAGCTGCCTCTGCCTTCCGAAAAAAAAGGAGGG 1844
AZM4_68974   1785 AAGAAGAAGAAAATGAGCAGCTGCCTCTGCCTTCCGAAAAAAAAGGAGGG 1834
                  *********** **************************************

                                3a/16a              11/16b
CA2 gene     1845 GCCAGCGAAGGAGAAGCCGTCCACAGATACCCCCACCTCGTCACTCCTTC 1894
AZM4_68974   1835 GCCAGCGAAGGAGAAGCCGTCCACAGATACCCCCACCTCGTCACTCCTTC 1884
                  **************************************************

                           16b               11
CA2 gene     1895 AGAACCAGAAGCCCTCC-AACCTCCACCTCCTCCCTCCAAGGCTTCCTCC 1943
AZM4_68974   1885 AGAACCAGAAGCCCTCCCAACCTCCACCTCCTCCCTCCAAGGCTTCCTCC 1934
                  ***************** ********************************

                 Exon2  end 16a/12
CA2 gene     1944 AAGGGC<CGTCCCCTCCTCCTCCTCCTCATCTTCCTCTCTCACCTTCAGCA 1993
AZM4_68974   1935 AAGGTC<CGTCCCCTCCTCCTCCTCCTCATCTTCCTCTCTCACCTTCAGCA 1984
                  **** *********************************************

                                                       3a
CA2 gene     1994 CCATCCTCCACACAGCAGCACGCGCGCAGCAATCTCACCGTTTTCTTTTC 2043
AZM4_68974   1985 CCATCCTCCACACAGCAGCACGCGCGCAGCAATCTCACCGTTTTCTTTTC 2034
                  **************************************************

CA2 gene     2044 CTCCATTGCCATCAGTAGCTAGCCACACTGCATGCATTCAGCTTCCGCTT 2093
AZM4_68974   2035 CTCCATTGCCATCAGTAGCTAGCCACACTGCATGCATTCAGCTTCCGCTT 2084
                  **************************************************
```

```
                                               3a
CA2 gene     2094 TCTCCCTGTGTAGCGAGCGCTGGTGCCGGCCGGTGCAGAGAAGATCCCTG 2143
AZM4_68974   2085 TCTCCCTGTGTAGCGAGCGCTG-TGCCGGCCGGTGCAGAG           2123
                  ********************** *****************


CA2 gene     2144 CTCCCCCCCCCCCCCCCCCCCCCTAATTAAGATCACCTTTGTGCATTTTT 2193
AZM4_68974   2124                                                    2123


                        4b                      3a
CA2 gene     2194 TTCCTTGTGTTGTGGTCCGTCGGCAAGTAGGCCAAAATTGCATCATGCCA 2243
AZM4_68974   2124                                                    2123


CA2 gene     2244 TGGCCCCTCCTCTTCTACTACCTCGTCATGCAGCCAGCAACGACATGAAT 2293
AZM4_68974   2124                                                    2123


                          4b
CA2 gene     2294 GACCCGAACGAAGTATCTGGCGTTGACATTGCAG 2327
AZM4_68974   2124                                    2123
```

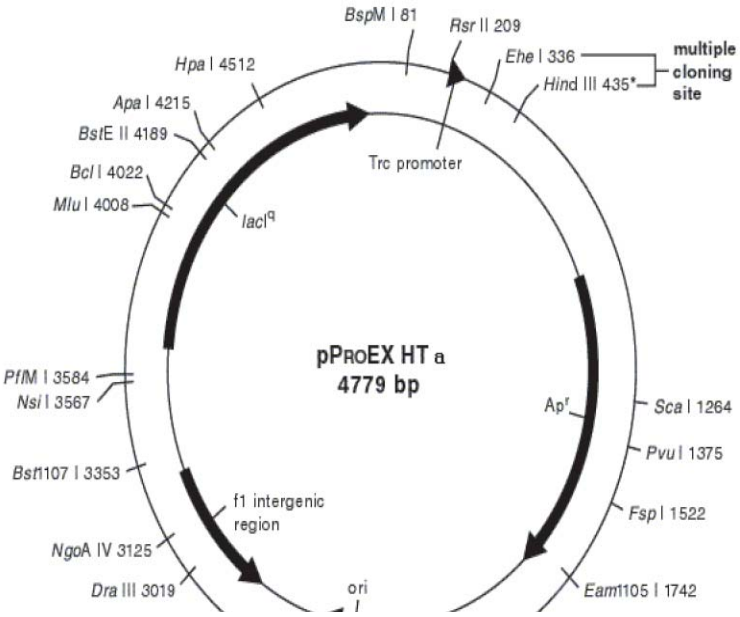
Appendix – Chapter 5

5.1: Vector diagrams

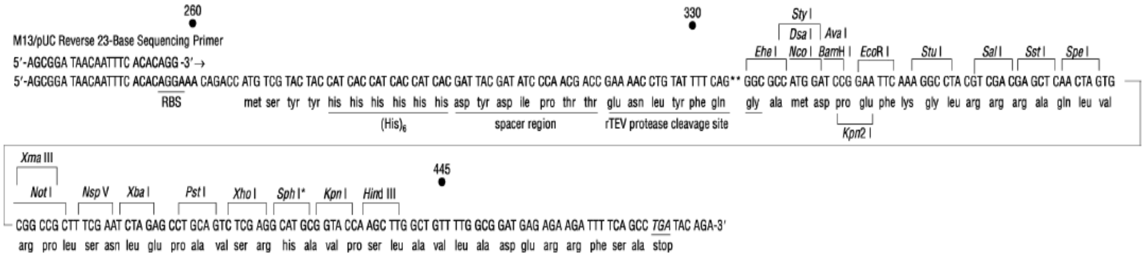
A. pGEM®-T (Promega)


Xmn I 1994
Sca I 1875
Nae I 2695
f1 ori
Amp^r
pGEM®-T Vector (3003bp)
lacZ
T
T
T7 ↓
1 start
Apa I 14
Aat II 20
Sph I 26
BstZ I 31
Nco I 37
Sac II 46
Spe I 55
Not I 62
BstZ I 62
Pst I 73
Sal I 75
Nde I 82
T7 Transcription Start
5′ . . . TGTAA TACGA CTCAC TATAG GGCGA ATTGG GCCCG ACGTC GCATG CTCCC GGCCG
3′ . . . ACATT ATGCT GAGTG ATATC CCGCT TAACC CGGGC TGCAG CGTAC GAGGG CCGGC
T7 Promoter
Apa I
Aat II
Sph I
BstZ I
CCATG GCCGC GGGATT3′ (cloned insert) ATCAC TAGTG CGGCC GCCTG CAGGT CGACC ATATG
GGTAC CGGCG CCCTA 3′TTAGTG ATCAC GCCGG CGGAC GTCCA GCTGG TATAC
Nco I
Sac II
Spe I
Not I
BstZ I
Pst I
Sal I
Nde I
SP6 Transcription Start
GGAGA GCTCC CAACG CGTTG GATGC ATAGC TTGAG TATTC TATAG TGTCA CCTAA AT . . . 3′
CCTCT CGAGG GTTGC GCAAC CTACG TATCG AACTC ATAAG ATATC ACAGT GGATT TA . . . 5′

B. pROEx^{TM}-HTa (Invitrogen)

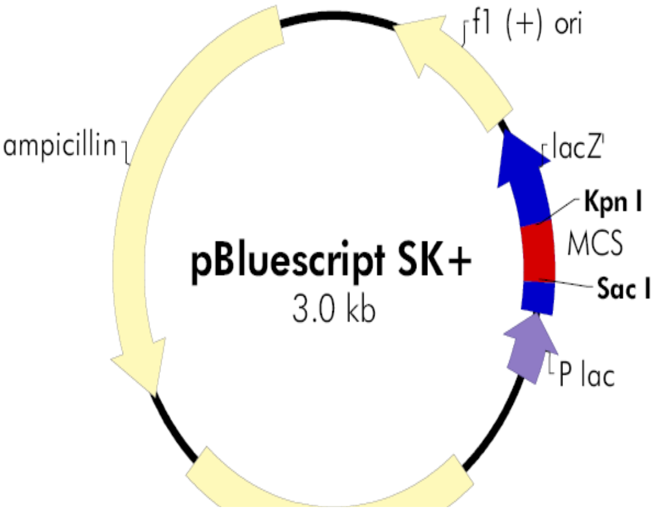


pPROEX HTa multiple cloning site and primer binding region: 235-482

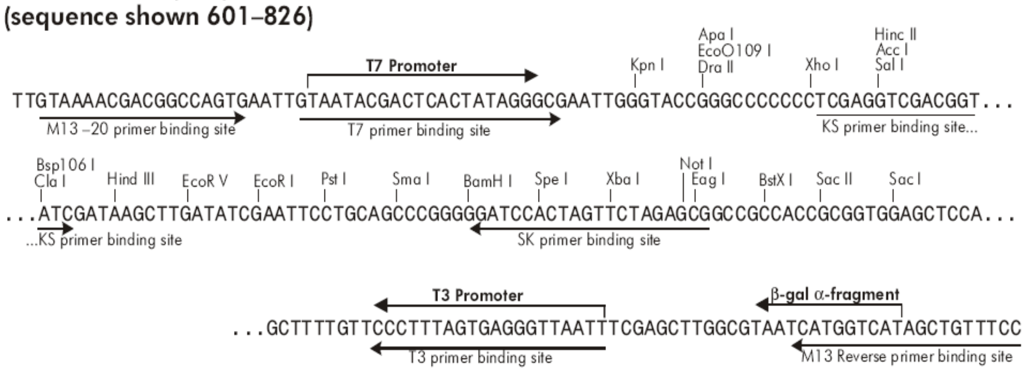

260
330
M13/pUC Reverse 23-Base Sequencing Primer
5′-AGCGGA TAACAATTTC ACACAGG -3′→
5′-AGCGGA TAACAATTTC ACACAGGAAA CAGACC ATG TCG TAC TAC CAT CAC CAT CAC CAT CAC GAT TAC GAT ATC CCA ACG ACC GAA AAC CTG TAT TTT CAG** GGC GCC ATG GAT CCG GAA TTC AAA GGC CTA CGT CGA CGA GCT CAA CTA GTG
RBS
met ser tyr tyr his his his his his his asp tyr asp ile pro thr thr glu asn leu tyr phe gln gly ala met asp pro glu phe lys gly leu arg arg arg ala gln leu val
(His)6
spacer region
rTEV protease cleavage site
Sty I
Dsa I
Ava I
Ehe I
Nco I
BamH I
EcoR I
Stu I
Sal I
Sst I
Spe I
Kpn2 I
Xma III
Not I
Nsp V
Xba I
Pst I
Xho I
Sph I*
Kpn I
Hind III
445
CGG CCG CTT TCG AAT CTA GAG CCT GCA GTC TCG AGG CAT GCG GTA CCA AGC TTG GCT GTT TTG GCG GAT GAG AGA AGA TTT TCA GCC TGA TAC AGA-3′
arg pro leu ser asn leu glu pro ala val ser arg his ala val pro ser leu ala val leu ala asp glu arg arg phe ser ala stop

*Sph I has 2 recognition sites in pPROEX HT.

C. pBluescript (Stratagene)



pBluescript SK (+/–) Multiple Cloning Site Region
(sequence shown 601–826)


T7 Promoter
Kpn I
Apa I
EcoO109 I
Dra II
Xho I
Hinc II
Acc I
Sal I
TTGTAAAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGAATTGGGTACCGGGCCCCCCCTCGAGGTCGACGGT...
M13 –20 primer binding site
T7 primer binding site
KS primer binding site...
Bsp106 I
Cla I
Hind III
EcoR V
EcoR I
Pst I
Sma I
BamH I
Spe I
Xba I
Not I
Eag I
BstX I
Sac II
Sac I
...ATCGATAAGCTTGATATCGAATTCCTGCAGCCCGGGGGATCCACTAGTTCTAGAGCGGCCGCCACCGCGGTGGAGCTCCA...
...KS primer binding site
SK primer binding site
T3 Promoter
β-gal α-fragment
...GCTTTTGTTCCCTTTAGTGAGGGTTAATTTCGAGCTTGGCGTAATCATGGTCATAGCTGTTTCC
T3 primer binding site
M13 Reverse primer binding site

5.2: Amino acid sequences of the expressed CA isoforms

For each of the sequences presented below, amino acids in italics are part of the histidine tag that is generated by the expression vector pROEx$^{\text{TM}}$. The terminating codon is indicated with an asterisk.

CA1

*MSYYHHHHHHDYDIPTTENLYFQGAMGS*MYTLPVRATTSSIVASLATPAPSSSSGSGRPRLRLIRNAPV
FAAPATVVGMDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVCPSVTLGLQPG
EAFTVRNIASMVPPYDKIKYAGTGSAIEYAVCALKVQVIVVIGHSCCGGIRALLSLKDGAPDNFHFVEDW
VRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEGLAGGTLKLVGAHYDFVKGQFV
TWEPPQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMVFACSDSRVCPSVTLGLQPGEAFTV
RNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIRALLSLQDGAPDTFHFVEDWVKIGFI
AKMKVKKEHASVPFDDQCSILEKEAVNVSLENLKTYPFVKEGLANGTLKLIGAHYDFVSGEFLTWKK*

Repeat A

*MSYYHHHHHHDYDIPTTENLYFQGAMGS*MYTLPVRATTSSIVASLATPAPSSSSGSGRPRLRLIRNAPV
FAASATVVGMDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVCPSVTLGLQPG
EAFTVRNIASMVPPYDKIKYAGTGSAIEYAVCALKVQVIVVIGHSCCGGIRALLSLKDGAPDNFHFVEDW
VRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEGLAGGTLKLVGAHYDFVKGQFV
TWEPP*

Repeat B

*MSYYHHHHHHDYDIPTTENLYFQGA*MDPPQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMV
FACSDSRVCPSVTLGLQPGEAFTVRNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIR
ALLSLKDGAPDNFHFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEGLA
GGTLKLVGAHYDFVKGQFVTWEPPQDALEACGTKLGCFGG*

Repeat C

*MSYYHHHHHHDYDIPTTENLYFQGS*MDPPQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMV
FACSDSRVCPSVTLGLQPGEAFTVRNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIR
ALLSLQDGAPDTFHFVEDWVKIGFIAKMKVKKEHASVPFDDQCSILEKEAVNVSLENLKTYPFVKEGLAN
GTLKLIGAHYDFVSGEFLTWKK*

CA4

*MSYYHHHHHHDYDIPTTENLYFQGAMGS*MYTLPVRATTSSIVASLATPAPSSSSGSGRPRLRLIRNAPV
FAAPATVVGMDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVCPSVTLGLQPG
EAFTVRNIASMVPPYDKIKYAGTGSAIEYAVCALKVQVIVVIGHSCCGGIRALLSLKDGAPDNFHFVEDW
VRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEGLAGGTLKLVGAHYDFVKGQFV
TWEPPQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMVFACSDSRVCPSVTLGLQPGEAFTV
RNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIRALLSLKDGAPDNFHFVEDWVRIGS
PAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEGLAGGTLKLVGAHYDFVKGQFVTWEP
P*

CA5

*MSYYHHHHHHDYDIPTTENLYFQGA*MDPPQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMVFACSDSRVCPSVTLGLQPGEAFTVRNIAAMVPGYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIRALLSLKDGAPDNFHFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEGLAGGTLKLVGAHYDFVKGQFVTWEP*

CA6

*MSYYHHHHHHDYDIPTTENLYFQGA*MVPPYDKIKYAGTGSAIEYAVCALKVQVIVVIGHSCCGGIRALLSLKDGAPDNFHFVEDWVRIGSPAKNKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEGLAGGTLKLVGAHYDFVKGQFVTWEPPQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMVFACSDSRVCPSVTLGLQPGEAFTVRNIAAMVPGYDKTKYTGIGSAIEYAACALKVEVLVVIGHSCCGGIRALLSLQDGAPDTFHFVEDWVKIGFIATMKVKKEHASVPFDDQCSILEKEAVNVSLENLKTYPFVKEGLANGTLKLIGGHYDFVSGEFLTWKK*

CA7

*MSYYHHHPHQGYDIPTTENLYFQGA*MDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVCPSVTLGLQPGEAFTVRNIASMVPGYDKTKYTGIGSAIEYAACALKVEVLVVIGHSCCGGIRALLSLQDGAPDTFHFVEDWVKIGFIATMKVKKEHASVPFDDQCSILEKEAVNVSLENLKTYPFVKEGLANGTLKLIGGHYDFVSGEFLTWKK*

CA8

*MSYYHHHHHHDYDIPTTENLYFQGA*MDPTVERLKSGFQKFKTEVYDKKPELFEPLKSGQSPRYMVFACSDSRVCPSVTLGLQPGEAFTVRNIASMVPGYDKTKYTGIGSAIEYAACALKVEVLVVIGHSCCGGIRALLSLQDGAPDTFHFVEDWVKIGFIATMKVKKEHASVPFDDQCSILEKEAVNVSLENLKTYPFVKEGLANGTLKLIGGHYDFVSGEFLTWKK*

Appendix – Chapter 6

6.1: Deduced amino acid sequence of Exon 1 of the CA2 gene

MYTLPVRATTSSIVASLATPAPSSSSGSGRPRLRLIRNAPVFAAPATV

6.2: Deduced amino acid sequence of Repeat B

MPQDAIERLTSGFQQFKVNVYDKKPELFGPLKSGQAPKYMVFACSDSRVCPSVTLGLQPGEAFTVRNI
AAMVPGYDKTKYTGIGSAIEYAVCALKVEVLVVIGHSCCGGIRALLSLKDGAPDNFHFVEDWVRIGSPAK
NKVKKEHASVPFDDQCSILEKEAVNVSLQNLKSYPFVKEGLAGGTLKLVGAHYDFVKGQFVTWEP*