Biologia, Bratislava, **57/Suppl. 11:** 137—148, 2002

Characterisation of putative α-amylases from apple (*Malus domestica*) and Arabidopsis thaliana

Duncan Stanley[1,2], Anna M. Fitzgerald[2], Kevin J. F. Farnden[1] & Elspeth A. MacRae[2]*

[1]*Department of Biochemistry, University of Otago, P.O. Box 56, Dunedin, New Zealand*
[2]*HortResearch, Mt Albert Research Centre, Private Bag 92169, Auckland, New Zealand; tel.: ++ 649 815 4200, fax: ++ 649 815 4202, e-mail: EMacRae@hortresearch.co.nz*

Stanley, D., Fitzgerald, A. M., Farnden, K. J. F. & MacRae, E. A., Characterisation of putative α-amylases from apple (*Malus domestica*) and *Arabidopsis thaliana*. Biologia, Bratislava, **57/Suppl. 11:** 137—148, 2002; ISSN 0006-3088.

We have identified a number of new plant α-amylase-like genes from apple and *Arabidopsis thaliana*, two of which appear to be completely novel. The putative products of these novel genes are characterised by a large (460–480 amino acid) N-terminal extension, which includes a putative plastid-targeting peptide. Phylogenetic analysis of these and other α-amylase-like genes from plants suggests that there are three distinct families of α-amylases in plants. These three families can also be separated on the basis of their intron structure, the relative size of domain B, and their predicted subcellular targeting. All three of the gene families appear to have the catalytic residues necessary for α-amylase activity, although it is possible that each may have different substrate specificity. Furthermore, each gene family appears to be targeted to different compartments within the cell. ESTs corresponding to each of the families have been identified in the gymnosperm *Pinus taeda*, suggesting that the three gene families diverged from each other prior to the angiosperm/gymnosperm split.

Key words: α-amylase, *Arabidopsis thaliana*, intron structure, *Malus domestica*, phylogeny, subcellular localisation.

Introduction

Starch is a vital carbon storage molecule for most plants. It is insoluble and stored within the plastids of plant cells. Starch can be separated into one of two basic groups by its function; transitory (or diurnal) starch is produced in photosynthetic tissues during the day, and at night it is broken down into sugars and exported to other parts of the plant. The recipient tissues can feed the sugars into many metabolic pathways, one of which is resynthesis of starch for long-term storage. Storage starch is produced in a variety of organs, including roots, tubers, bark and developing fruit.

α-Amylases (E.C. 3.2.1.1) are starch endohydrolases, which cleave α-1,4-glucan bonds within starch molecules. They have three domains: domain A folds into a $(\beta/\alpha)_8$ barrel, and contains the catalytic residues of the enzyme; domain B is a large loop that protrudes from between

* Corresponding author

the third β-strand and the third α-helix of domain A; domain C is located at the C-terminal end of the $(\beta/\alpha)_8$ barrel, and is made up of β-strands. The functions of domains B and C are not clear, although domain B has been shown to influence several isozyme-specific properties of barley α-amylases, including substrate binding, catalysis, and stability at low pH (RODENBURG et al., 1994; JUGE et al., 1995). The crystal structure of barley α-amylase AMY2-2 shows three Ca^{2+} binding sites in domain B and calcium ions are thought to be crucial for maintaining the domain's structure (KADZIOLA et al., 1994).

Much of the work on plant α-amylases has focussed on enzymes from monocotyledonous plants, particularly the cereals barley and rice. α-Amylase plays a vital role in the germination of cereal grains; it is secreted from the aleurone layer into the endosperm where it initiates starch hydrolysis (BECK & ZIEGLER, 1989). However α-amylases are not limited to tissues of germinating seeds. α-Amylase activity has also been detected in water-stressed leaves of barley (JACOBSEN et al., 1986), and in cultured rice cells, localised to the cell wall and amyloplasts (CHEN et al., 1994). Molecular studies of α-amylase genes in monocots have revealed a large but well-conserved family, often represented by multiple genes in each plant, e.g. at least 10 genes in rice (RANJIHAN et al., 1991). The proteins encoded by these genes all have N-terminal signal peptides, which would direct them into the cellular secretory pathway (CHAN et al., 1994).

Dicot α-amylases have received very little attention compared to the monocot enzymes, possibly because they are not as directly involved in industrial processes (e.g. brewing). Amylases have been implicated in several stress responses, including pathogen infection (HEITZ et al., 1991), and low temperature (HILL et al., 1996; WEGRZYN et al., 2000). A limited number of α-amylase genes have been identified in dicots, predominantly from expression in germinating cotyledons (YAMAUCHI et al., 1994). These α-amylases have high homology to the previously-characterised monocot genes and include putative signal peptides. However two genes, one from apple and a second from potato, have slightly shorter B domains and lack the usual signal peptide (WEGRZYN et al., 2000). These genes have been shown to form a separate phylogenetic clade, distinct from other plant α-amylases (WEGRZYN et al., 2000; PUJADAS & PALAU, 2001).

The completion of the *Arabidopsis thaliana* genome, together with continuing sequencing projects in rice, loblolly pine and other model plants, has made public a huge amount of sequence information. We have also had access to a fruit tree EST database, at HortResearch, New Zealand, which includes sequences from apple (*Malus domestica*) and kiwifruit (*Actinidia deliciosa*). Sequence data from these sources was examined for possible α-amylase-encoding genes, focussing on apple and *Arabidopsis*.

Material and methods

Mature apple fruit (*Malus domestica* Borkh. cv. Granny Smith) were harvested and stored at 4 ℃ for 8 days, then frozen in liquid nitrogen and used for RNA extraction, as in LANGENKAMPER et al. (1998). RT-PCR was performed, using Superscript II (Invitrogen), followed by Expand Hi-Fidelity PCR enzyme (Roche Molecular Diagnostics). PCR of genomic DNA used only Expand Hi-Fidelity. PCR fragments were cloned into pGEMT-Easy (Promega) prior to sequencing. Sequences were assembled into contigs using the Fragment Assembly System of the GCG Wisconsin Package (DEVEREUX et al., 1984).

Screening of Genbank and the HortResearch EST database was performed using the BLAST tool (ALTSCHUL et al., 1997) at the NCBI and HortResearch. Intron/exon splicing predictions of *Arabidopsis* genes were performed by Eukaryotic GeneMark.hmm (LUKASHIN & BORODOVSKY, 1998), located at: `http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi`.

α-Amylase protein sequences for alignments were extracted from the Genbank database. The first two letters of each sequence name denote the source: At = *Arabidopsis thaliana*, Cr = *Cuscuta reflexa* (southern Asian dodder), Hv = *Hordeum vulgare* (barley), Md = *Malus domestica* (apple), Os = *Oryza sativa* (rice), Pv = *Phaseolus vulgaris* (French bean), St = *Solanum tuberosum* (potato), Ta = *Triticum aestivum* (wheat), Vm = *Vigna mungo* (black gram), Zm = *Zea mays* (maize). The following sequences are referred to in text, Table 1, and Figures 3, 4; accessions are for Genbank protein records: Cramy1 (AAA16513), Hvamy2-2 (AAA98790), Hvamy1 (AAA32925), HvamyA (AAA32929), Hvamyaa24 (CAA72143), Mdamy8 (AAF63239), OsamyA (AAA33885), OsamyC2M (CAA45903), Osamy3B (AAA33897), OsamyB (AAA33886), Pvamy1 (BAA33879), Stamy21 (AAA91883), Stamy23 (AAA91884), Taamy3 (AAA34259), Vmamy1 (CAA37217), Zmamy1 (AAA50161).

For functional comparisons, amino acid sequence alignments were performed using the program Pileup, part of the GCG Wisconsin Package (DEVEREUX et al., 1984), and shaded using the program BOXSHADE.

For the phylogenetic analysis, amino acid sequence alignments were performed using the program CLUSTALW (THOMPSON et al., 1994). The resulting alignments were used to calculate 'protein distances' by the Dayhoff method, which were used to

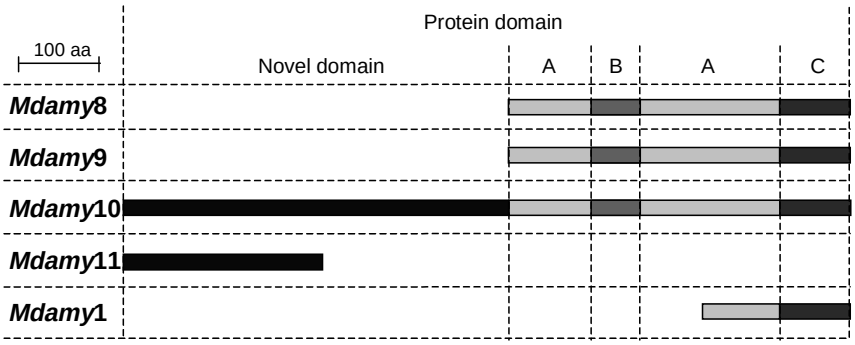


Fig. 1. Comparative alignment of the available sequence data from apple α-amylase genes. Each bar is shaded to distinguish the four protein domains encoded by the sequences: A, B and C represent the structural domains A, B and C, which are found throughout all α-amylases. Novel domain refers to the N-terminal domain found only in *Mdamy*10, *Mdamy*11 and *Atamy*1. *Mdamy*1 and *Mdamy*11 are not full-length cDNA clones.

construct a phylogenetic tree by the neighbour-joining method. Both functions were performed using PHYLIP version 3.5c (Felsenstein, 1995). Bootstrap analysis (Felsenstein, 1995) was performed with 1000 bootstrap replicates. Trees were visualised using TreeView (Page, 1996).

Targeting motifs were predicted using the servers at the Center for Biological Sequence Analysis (`http://www.cbs.dtu.dk/services/`). TargetP (Emanuelsson et al., 2000) was used for initial predictions, followed by SignalP (Nielsen et al., 1997) and ChloroP (Emanuelsson et al., 1999), for more in-depth analysis of putative signal peptides and plastid-targeting peptides, respectively.

Results

Identification of α-amylase-like genes from Apple
In previous work Wegrzyn et al. (2000) identified an α-amylase-like cDNA from apple fruit, termed *Mdamy*8. The *Mdamy*8 gene has now been amplified and sequenced from genomic DNA. We have identified a second *Mdamy*8-like α-amylase gene in apple (*Mdamy*9) from a pair of 5' and 3' RACE products, originally thought to be from *Mdamy*8. Genomic sequence amplified from within the range of the RACE products had revealed that the intron sequences of *Mdamy*9 were significantly different from those of the *Mdamy*8 gene (data not shown). RT-PCR was performed with apple RNA using a forward primer designed to the 5' RACE sequence and a reverse primer designed to the 3' RACE sequence. The same primers were also used to amplify the *Mdamy*9 gene from genomic DNA. *Mdamy*8 and 9 are 94% identical at the amino acid level, but their 5' and 3' untranslated regions (UTRs) are only 36% identical at the nucleotide level. This is also reflected in the introns of the two genes, which correlate in size and position but are significantly different in sequence. The HortResearch EST database contains two sequences from the *Mdamy*8 gene, both found in an EST library made from senescing apple leaf tissue.

Another pair of sequences from the same RNA library show high similarity (78% at the amino acid level) to *amyVm*1 from *Vigna mungo* (Yamauchi et al., 1994) (referred to as *Vmamy1* in Figure 4). Further sequencing of the cDNAs has created a sequence contig 700 bp long, including 205 bp of 3' UTR sequence; the available coding sequence encodes 160 of a total expected 420 amino acids. This has been tentatively named *Mdamy*1 (Fig. 1).

However, the vast majority of α-amylase-like sequences discovered within the HortResearch database do not correspond to previously-characterised α-amylases. The transcripts are from a single gene, and are present in numerous tissue libraries, including floral bud, petal, young fruit and mature fruit libraries. Blast searches against the Genbank database revealed a match to a putative α-amylase from *Arabidopsis,* as well as to a number of ESTs from a variety of plants, including *Arabidopsis*, *Lycopersicon esculentum* (tomato), *Medicago truncatula, Glycine max* (soybean), *Solanum tuberosum* (potato) and *Pinus taeda* (loblolly pine). The full-length cDNA sequence, named *Mdamy*10, was obtained by a mixture of 5' RACE (using apple RNA from floral buds), and identification and analysis of overlapping EST sequences. The coding region of *Mdamy*10 is 2706 bp long, encoding 901 amino acids, the last 400 amino acids of which show modest similarity to both *Mdamy*8 (54%) and

*amyVm*1 (58%); however the first 500 amino acids are not found in any characterised α-amylase (Fig. 1).

The only apple EST sequence that matched the 5' region of *Mdamy*10 was from a full length cDNA, displaying 79% nucleotide identity, but only 900 nucleotides in length (Fig. 1). The transcript appears to be the product of a gene homologous to *Mdamy*10, except that the equivalent of the third intron of *Mdamy*10 has not been spliced from the transcript. This results in cleavage and polyadenylation of the transcript within the third intron, and potential production of a truncated protein. We have tentatively labelled the transcript *Mdamy*11, despite the lack of an α-amylase domain, in anticipation of recovering a full-length cDNA or genomic sequence encoding all domains.

Identification of α-amylase-like genes from Arabidopsis

Three *Arabidopsis* genes, which appear to encode α-amylases, have been identified by their homology to genes from apple and *V. mungo*. The gene *Atamy*1 is related to *amyVm*1, showing 68% identity and 73% similarity at the amino acid level, and is found on chromosome 4 (BAC F13M23). *Atamy*2 is a homologue of *Mdamy*8, and displays 80% identity and 85% similarity at the amino acid level; it is located on chromosome 1 (BAC T23E18). Neither gene has been correctly annotated in Genbank; both have incorrect prediction of intron splice sites.

*Atamy*3 was originally annotated in Genbank by TIGR, predicted from genomic sequence. The annotation suggested a coding sequence 3308bp in length, encoding a protein of 826 amino acids (compared to around 415 aa for other plant α-amylases). We reanalysed the appropriate chromosomal segment, using Eukaryotic GeneMark.hmm, (LUKASHIN & BORODOVSKY, 1998) to predict gene splicing. This algorithm predicted an even larger coding region, due to inclusion of two extra exons, producing a protein 887 aa in size, displaying 68 % identity to the predicted *Mdamy*10 product. This sequence prediction has been confirmed from the sequencing of an *Atamy*3 cDNA clone at the Salk Institute, CA., published in Genbank in August 2001 (accession AY050398). *Atamy*3 is also located on chromosome 1 (BAC T17F3), approximately 2.2 Mb away from *Atamy*2. Blast searches have not yielded any other putative α-amylase genes present in the *Arabidopsis* genome.

Structure of α-amylase and α-amylase-like genes in plants

The gene structures of plant α-amylases have been compared previously, most significantly in HUANG et al. (1990). Repeating this type of analysis, but using new *Arabidopsis* and apple α-amylase sequences, produces a far more complicated picture of α-amylase intron evolution. Previously-sequenced α-amylase genes have at most three introns. Some cereal genes appear to have lost the second intron (HUANG et al., 1990), but the position of each intron is conserved between genes and species. The *Arabidopsis* homologue of cereal α-amylases, *Atamy*1, contains a fourth intron, not seen before in plant α-amylase genes, situated between the second and the last introns (Tab. 1).

The intron/exon structure of *Mdamy*8 and *Mdamy*9 is significantly different from previously-sequenced α-amylase genes (Tab. 1). Both genes contain 12 introns, one of which is found in the 5' UTR, and 13 exons, compared to other characterised α-amylases, which have from 2 to 4 introns. *Mdamy*8 and *Mdamy*9 do share some intron boundaries with other α-amylases, for example introns 4 and 5 correspond to introns 1 and 2 of *Atamy*1, but they do not have introns in the same position as the third and fourth introns of *Atamy*1. *Atamy*2 appears to have the same intron structure as *Mdamy*8 and *Mdamy*9; a putative intron has been located upstream of the translation start site, but it is not known whether the intron is spliced from the primary transcript (indicated by +/- in Table 1). Overall, the introns of *Atamy*2 are shorter than those of *Mdamy*8 and *Mdamy*9; the 11 introns within the coding sequence of *Atamy*2 total 1026 bp in size; the equivalent introns of *Mdamy*8 total 2143 bp.

Of the 12 introns that interrupt the coding sequence of *Atamy*3, the last 6 are within the amylase-encoding region of the gene (i.e. the 3' half). *Atamy*3 has no intron boundaries in common with *Atamy*1, however the tenth and eleventh introns of *Atamy3* do correlate to the ninth and tenth introns of *Mdamy*8 (Tab. 1).

Post-transcriptional processing of α-amylase-like genes

Initial sequencing of *Mdamy*8 transcripts revealed up to five different 5' RACE products, which fell into three groups of 5' UTR sequence (WEGRZYN et al., 2000); one of these (labelled as 'E' in the original paper) has since been identified as a transcript from the *Mdamy*9 gene. Further PCR of the 5' UTR, using genomic DNA and cDNA, has revealed that *Mdamy*8 produces at least 3 different

Table 1. The distribution of introns in α-amylase genes.[a]

	1	2	3	4	5	6	7	8	9
	UTR	8–0	19–1	32–0	76–1	89–2	159–0	161–0	194–2
Hvamy2-2	–	–	–	+	–	–	–	–	–
*Vmamy*1	–	–	–	+	+	–	–	–	–
*Atamy*1	–	–	–	+	+	–	–	–	–
*Mdamy*8	+	+	+	+	+	–	–	+	–
*Mdamy*9	+	+	+	+	+	–	–	+	–
*Atamy*2	+/–	+	+	+	+	–	–	+	–
*Atamy*3	–	–	–	–	–	+	+	–	+

	10	11	12	13	14	15	16	17	18
	209–2	221–0	248–1	280–0	322–0	334–0	348–0	367–0	381–0
Hvamy2-2	–	–	–	–	–	–	+	–	–
*Vmamy*1	–	–	–	–	–	–	+	–	–
*Atamy*1	–	+	–	–	–	–	+	–	–
*Mdamy*8	+	–	+	+	+	+	–	+	–
*Mdamy*9	+	–	+	+	+	+	–	+	–
*Atamy*2	+	–	+	+	+	+	–	+	–
*Atamy*3	–	–	–	+	+	–	–	–	+

[a] The presence of an intron is represented by a + sign and its absence by a – sign; +/– indicates a putative intron. The 5' portion of *Atamy*3 (codons 1-470) has not been included in the analysis. We have defined 18 distinct intron positions, numbered from the 5' of the gene sequences and defined by codon number and phase; the codon numbers refer to the amino acid alignment in Figure 3. For introns separating codon after the first base, the codon number is followed by −1. For introns separating codon after the second base, the codon number is followed by −2. For introns falling between codons, the number of the codon located 3' of the intron is given, followed by −0.

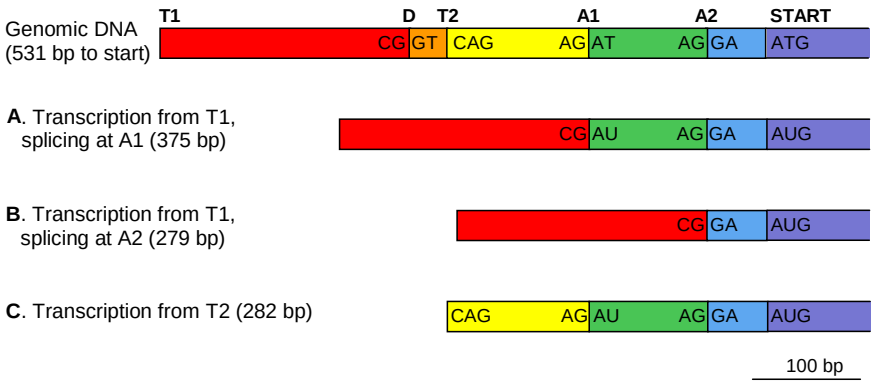
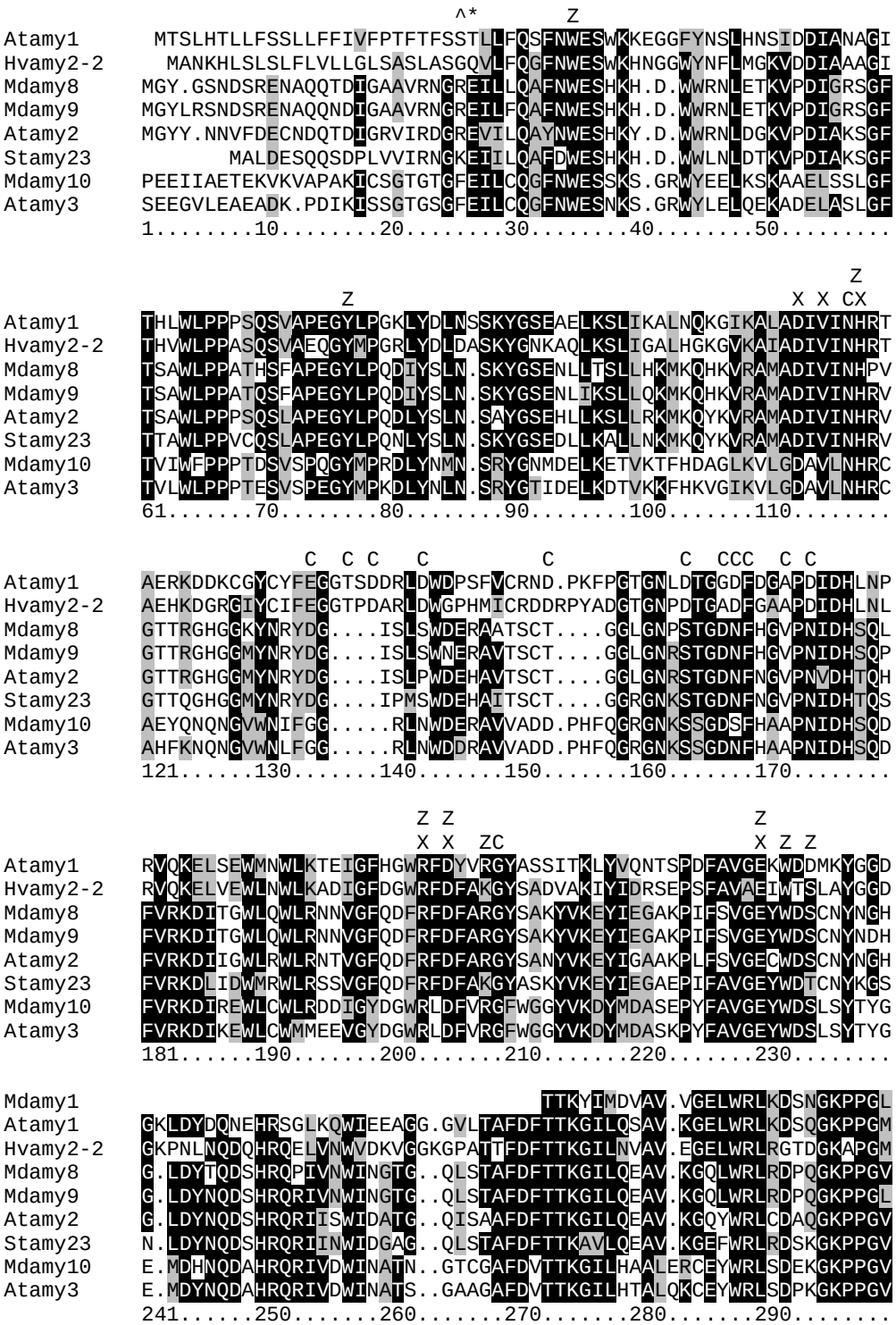


Fig. 2. Diagram of the alternate splicing/transcription events in the 5' UTR of *Mdamy*8; boxes represent DNA/RNA. Important sequence elements of the UTR are indicated by letters above the genomic DNA box; T1 = primer position 69bp downstream of a putative transcription start site; T2 = putative transcription start site within intron sequence; D = intron donor splice site; A1 & A2 = alternate intron acceptor splice sites; START = translation start site. Sequence surrounding intron splice sites and transcription/translation start sites is shown within each box.

transcripts; the variation occurs around the first intron of the gene, located in the 5' UTR. Figure 2 shows the 3 transcripts compared to genomic DNA. T1 is not the actual transcription start site of products A and B, but rather the position of a PCR primer located 69 bp downstream of the


^* Z
Atamy1 MTSLHTLLFSSLLFFIVFPTFTFSSTLLFQSFNWESWKKEGGFYNSLHNSIDDIANAGI
Hvamy2-2 MANKHLSLSLFLVLLGLSASLASGQVLFQGFNWESWKHNGGWYNFLMGKVDDIAAAGI
Mdamy8 MGY.GSNDSRENAQQTDIGAAVRNGREILLQAFNWESHKH.D.WWRNLETKVPDIGRSGF
Mdamy9 MGYLRSNDSRENAQQNDIGAAVRNGREILFQAFNWESHKH.D.WWRNLETKVPDIGRSGF
Atamy2 MGYY.NNVFDECNDQTDIGRVIRDGREVILQAYNWESHKY.D.WWRNLDGKVPDIAKSGF
Stamy23 MALDESQQSDPLVVIRNGKEIILQAFDWESHKH.D.WWLNLDTKVPDIAKSGF
Mdamy10 PEEIIAETEKVKVAPAKICSGTGTGFEILCQGFNWESSKS.GRWYEELKSKAAELSSLGF
Atamy3 SEEGVLEAEADK.PDIKISSGTGSGFEILCQGFNWESNKS.GRWYLELQEKADELASLGF
1.....10.....20.....30.....40.....50.....
Z
Z X X CX
Atamy1 THLWLPPPSQSVAPEGYLPGKLYDLNSSKYGSEAELKSLIKALNQKGIKALADIVINHRT
Hvamy2-2 THVWLPPASQSVAEQGYMPGRLYDLDASKYGNKAQLKSLIGALHGKGVKAIADIVINHRT
Mdamy8 TSAWLPPATHSFAPEGYLPQDIYSLN.SKYGSENLLTSLLHKMKQHKVRAMADIVINHPV
Mdamy9 TSAWLPPATQSFAPEGYLPQDIYSLN.SKYGSENLIKSLLQKMKQHKVRAMADIVINHRV
Atamy2 TSAWLPPPSQSLAPEGYLPQDLYSLN.SAYGSEHLLKSLLRKMKQYKVRAMADIVINHRV
Stamy23 TTAWLPPVCQSLAPEGYLPQNLYSLN.SKYGSEDLLKALLNKMKQYKVRAMADIVINHRV
Mdamy10 TVIWFPPPTDSVSPQGYMPRDLYNMN.SRYGNMDELKETVKTFHDAGLKVLGDAVLNHRC
Atamy3 TVLWLPPPTESVSPEGYMPKDLYNLN.SRYGTIDELKDTVKKFHKVGIKVLGDAVLNHRC
61.....70.....80.....90.....100.....110.....
C C C C C C CCC C C
Atamy1 AERKDDKCGYCYFEGGTSDDRLDWDPSFVCRND.PKFPGTGNLDTGGDFDGAPDIDHLNP
Hvamy2-2 AEHKDGRGIYCIFEGGTPDARLDWGPHMICRDDRPYADGTGNPDTGADFGAAPDIDHLNL
Mdamy8 GTTRGHGGKYNRYDG....ISLSWDERAATSCT....GGLGNPSTGDNFHGVPNIDHSQL
Mdamy9 GTTRGHGGMYNRYDG....ISLSWNERAVTSCT....GGLGNRSTGDNFHGVPNIDHSQP
Atamy2 GTTRGHGGMYNRYDG....ISLPWDEHAVTSCT....GGLGNRSTGDNFNGVPNVDHTQH
Stamy23 GTTQGHGGMYNRYDG....IPMSWDEHAITSCT....GGRGNKSTGDNFNGVPNIDHTQS
Mdamy10 AEYQNQNGVWNIFGG.....RLNWDERAVVADD.PHFQGRGNKSSGDSFHAAPNIDHSQD
Atamy3 AHFKNQNGVWNLFGG.....RLNWDDRAVVADD.PHFQGRGNKSSGDNFHAAPNIDHSQD
121.....130.....140.....150.....160.....170.....
Z Z Z
X X ZC X Z Z
Atamy1 RVQKELSEWMNWLKTEIGFHGWRFDYVRGYASSITKLYVQNTSPDFAVGEKWDDMKYGGD
Hvamy2-2 RVQKELVEWLNWLKADIGFDGWRFDFAKGYSADVAKIYIDRSEPSFAVAEIWTSLAYGGD
Mdamy8 FVRKDITGWLQWLRNNVGFQDFRFDFARGYSAKYVKEYIEGAKPIFSVGEYWDSCNYNGH
Mdamy9 FVRKDITGWLQWLRNNVGFQDFRFDFARGYSAKYVKEYIEGAKPIFSVGEYWDSCNYNDH
Atamy2 FVRKDIIGWLRWLRNTVGFQDFRFDFARGYSANYVKEYIGAAKPLFSVGECWDSCNYNGH
Stamy23 FVRKDLIDWMRWLRSSVGFQDFRFDFAKGYASKYVKEYIEGAEPIFAVGEYWDTCNYKGS
Mdamy10 FVRKDIREWLCWLRDDIGYDGWRLDFVRGFWGGYVKDYMDASEPYFAVGEYWDSLSYTYG
Atamy3 FVRKDIKEWLCWMEEVGYDGWRLDFVRGFWGGYVKDYMDASKPYFAVGEYWDSLSYTYG
181.....190.....200.....210.....220.....230.....
Mdamy1 TTKYIMDVAV.VGELWRLKDSNGKPPGL
Atamy1 GKLDYDQNEHRSGLKQWIEEAGG.GVLTAFDFTTKGILQSAV.KGELWRLKDSQGKPPGM
Hvamy2-2 GKPNLNQDQHRQELVNWVDKVGGKGPATTFDFTTKGILNVAV.EGELWRLRGTDGKAPGM
Mdamy8 G.LDYTQDSHRQPIVNWINGTG..QLSTAFDFTTKGILQEAV.KGQLWRLRDPQGKPPGV
Mdamy9 G.LDYNQDSHRQRIVNWINGTG..QLSTAFDFTTKGILQEAV.KGQLWRLRDPQGKPPGL
Atamy2 G.LDYNQDSHRQRIISWIDATG..QISAAFDFTTKGILQEAV.KGQYWRLCDAQGKPPGV
Stamy23 N.LDYNQDSHRQRIINWIDGAG..QLSTAFDFTTKAVLQEAV.KGEFWRLRDSKGKPPGV
Mdamy10 E.MDHNQDAHRQRIVDWINATN..GTCGAFDVTTKGILHAALERCEYWRLSDEKGKPPGV
Atamy3 E.MDYNQDAHRQRIVDWINATS..GAAGAFDVTTKGILHTALQKCEYWRLSDPKGKPPGV
241.....250.....260.....270.....280.....290.....

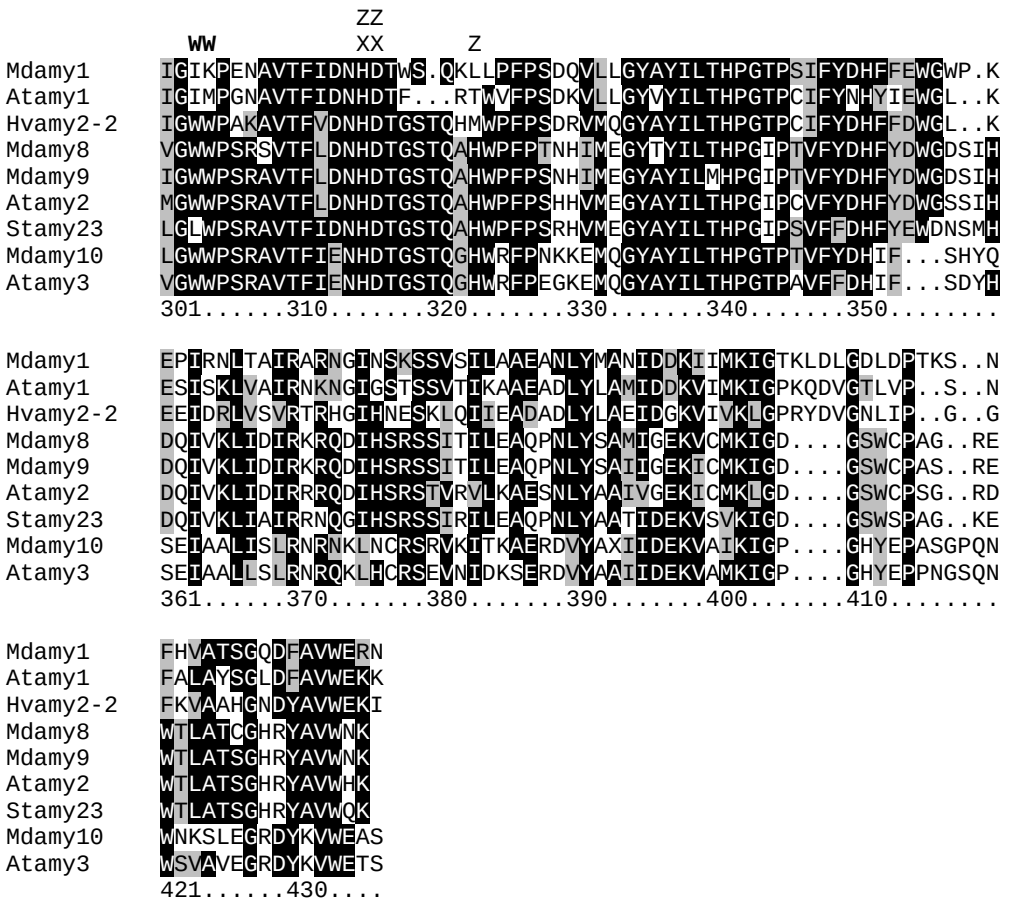


Fig. 3. PILEUP alignment of amino acid sequences from selected α-amylases (sequence details in Materials and methods). At = *Arabidopsis thaliana*, Hv = *Hordeum vulgare* (barley), Md = *Malus domestica* (apple), St = *Solanum tuberosum* (potato). Refer to Figure 4 for the family grouping of each sequence (i.e. Family one, two or three). Invariant residues conserved in starch hydrolase are marked X. Residues thought to be involved in Ca^{2+} binding in Hvamy2-2 are marked C. Z denotes residues of Hvamy2-2 involved in hydrogen bonding with the α-amylase inhibitor acarbose. The N-terminal amino acid of the mature barley enzyme is marked *; the predicted N-terminal amino acid of the mature Atamy1 protein is marked ^. The first 483 and 470 amino acids of *Mdamy*10 and *Atamy*3, respectively, have been removed.

putative start site of product B. We have not confirmed whether the same start site is used for A.

The 5' UTR intron can be spliced using the donor site, D, and either of two acceptor splice sites, A1 or A2, located 96 bp apart in *Mdamy*8 (Fig. 2), effectively allowing inclusion or removal of this 96 bp region from the messenger RNA. A third *Mdamy*8 transcript, product C, appears to initiate within the 5' UTR intron sequence, at position T2 (Fig. 2). However it is not clear whether this is due to a transcriptional initiation site within the intron or whether it is a partial product from an unspliced transcript.

The 5' UTR of *Mdamy*9 contains an intron in an equivalent position to that of *Mdamy*8, and also appears to have both acceptor sites intact, but in this case only one splice variant has been detected, corresponding to the full excision of the intron (the equivalent of product B of *Mdamy*8 – see Fig. 2).

Table 2 compares the untranslated regions of α-amylase transcripts from apple and *Arabidopsis*. The 5' UTRs of *Mdamy*8 and *Mdamy*9 are considerably longer than is normal for plant transcripts; up to 360 bp, whereas a typical plant transcript has 50–100 bp. *Mdamy*8 also has a very large 3' UTR, and at least two different polyadenylation sites in its 3' UTR (Wegrzyn et al., 2000). Only

Table 2. Characteristics of the untranslated regions (UTRs) of α-amylase genes from apple and *Arabidopsis*.

	5' UTR Length (Max/Min) (bp)	3' UTR Length (Max/Min) (bp)	Intron in 5'UTR?	No. of polyadenylation sites detected
*Mdamy*1	Unknown[a]	205	Unknown	1
*Atamy*1	Unknown	Unknown	Unknown	Unknown
*Mdamy*8	375 / 282	412 / 386	Yes	2
*Mdamy*9	305	120	Yes	1
*Atamy*2	Unknown	Unknown	Putative	Unknown
*Mdamy*10	19	557/428	No	4
*Mdamy*11	35	98	No	1
*Atamy*3	46	229	No	Unknown

[a] 'Unknown' indicates no sequence data available.

one polyadenylation site has been detected in the much shorter 3' UTR of *Mdamy*9.

Analysis of sequences upstream of the translation start sites of *Atamy*2, and the potato homologue, amy23 (GAUSING & KREIBERG, 1994), revealed possible intron acceptor and donor sites located in positions equivalent to those of *Mdamy*8/9 (data not shown). However, the size of the UTRs, whether the putative intron sequence is included in the primary transcript, and whether it is spliced or not, cannot be determined purely from these sequences.

In contrast to the *Mdamy*8 transcripts, *Mdamy*10, *Mdamy*11 and *Atamy*3 all appear to have short 5' UTRs – between 19 bp and 46 bp. However, *Mdamy*10 has a very long 3' UTR, up to 557 bp, although one of the four different polyadenylation sites can produce a 3' UTR as short as 428 bp. The 3' UTR sequence of *Atamy*3 is only 229 bp. The sizes of the *Atamy*1 and *Atamy*2 transcripts are unknown.

Conservation of α-amylase motifs

We have examined the new α-amylase sequences for residues implicated in binding calcium ions and catalysis. Figure 3 shows an amino acid alignment of the expected products of *Mdamy*1, *Mdamy*8, *Mdamy*9 and *Mdamy*10, plus all three genes from *A. thaliana* (the N-terminal portions of Mdamy10 and Atamy3 have been removed leaving only the α-amylase-like domains, A, B and C (Fig. 1)). We have also included the barley protein Hvamy2-2 (Genpept accession P04063), for which the crystal structure has been solved (KADZIOLA et al., 1994).

The alignment shows that variation at any particular residue is far more likely to be conserved between homologous genes than it is within a species. The three groups of genes can easily be sorted based upon the size of domain B: *Atamy*1 encodes the longest B domain of the new sequences, while in *Mdamy*8, *Mdamy*9 and *Atamy*2, domain B is 8 residues shorter than *Atamy*1, and *Mdamy*10 and *Atamy*3 are 6 amino acids shorter than *Atamy*1 (Fig. 3). In barley α-amylase AMY2-2, almost all of the amino acids involved in Ca^{2+} binding are found in domain B (marked by a C in Fig. 3). These critical residues are conserved in *Atamy*1, but most of the residues are either missing or altered in the other aligned sequences (Fig. 3).

The main active site residues of α-amylases appear to be intact in all of the new sequences, including two aspartates and a glutamate involved in the catalytic mechanism (positions 205, 316 and 230, respectively, of the alignment in Fig. 3) (SVENSSON, 1994). Most of the residues shown to bind acarbose in Hvamy2-2 (KADZIOLA et al., 1998) are also conserved (marked Z in Fig. 3)

Evolutionary organisation of plant α-amylases

We have performed a phylogenetic analysis of plant α-amylases, including the putative products of the genes presented in this paper and a selection of sequences from other plants. We have also assembled a set of overlapping ESTs from maize (*Zea mays*; accession numbers BE123285, AI782908 & AI999850) into a near full-length sequence, *Zmamy*2, missing only the first 25 codons of its sequence. The potato sequence Stamy21 is missing the first 75-80 amino acids. All other sequences are full length.

The phylogenetic tree (Fig. 4) shows that plant α-amylases are organised into three distinct families, with each family represented by a single gene in *Arabidopsis*. All of the major branches of the tree are well supported by bootstrap analysis. Family one, which is composed of almost all of the previously characterised α-amylases and has the most representatives sequenced, is clearly divided into monocot and dicot sequences.

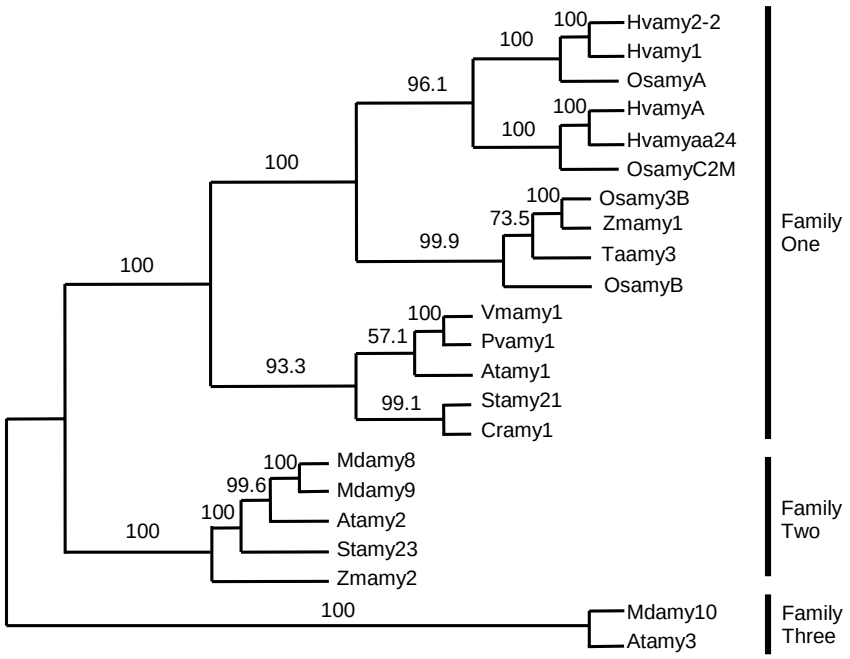


Fig. 4. Unrooted neighbour-joining tree constructed from Dayhoff distances from full-length and near-full-length protein sequences coding plant α-amylases (sequence details in Materials and methods). Abbreviations are: At = *Arabidopsis thaliana*, Cr = *Cuscuta reflexa* (southern Asian dodder), Hv = *Hordeum vulgare* (barley), Md = *Malus domestica* (apple), Os = *Oryza sativa* (rice), Pv = *Phaseolus vulgaris* (French bean), St = *Solanum tuberosum* (potato), Ta = *Triticum aestivum* (wheat), Vm = *Vigna mungo* (black gram), Zm = *Zea mays* (maize). The tree was created by the neighbour-joining method, bootstrap values are shown as percentages (1000 replicates).

In addition to the full-length sequences included in the phylogenetic analysis above, Genbank contains several partial sequences that are of interest in terms of molecular evolution. A single, partial EST from maize (BG842601) has surprisingly high homology to *Atamy*3, 89% identity over the 182 amino acids that it encodes. It displays much lower identity to other maize α-amylase sequences; 61% identity to the putative *Zmamy*2 product (Family two), and only 51% to the Family one representative, Zmamy1. Although this is the only Family three sequence so far detected in monocots, it suggests that Family three α-amylases are present in both monocots and dicots.

Partial EST sequences representing each of the three families have been identified from *P. taeda*, for example Genbank accessions AW290674 (Family one), AW290712 (Family two) and AI725250 (Family three). All three of these sequences encode regions encompassing domain B, and they can easily be separated based upon the relative size of that domain.

Subcellular localisation of plant α-amylases

The subcellular localisations of plant α-amylase proteins were predicted using servers at the Center for Biological Sequence Analysis (`http://www.cbs.dtu.dk/services/`). The program TargetP (Emanuelsson et al., 2000) was used initially to predict the targeting of all the *Arabidopsis* and apple proteins, except *Mdamy*1 (which is only a partial sequence). The Family two proteins do not appear to be targeted to any particular compartment of the cell and are most likely cytosolic. *Atamy*1 (Family one) was predicted to enter the secretory pathway; the program SignalP (Nielsen et al., 1997) placed the cleavage site for the putative signal peptide between residues 24 and 25 (Fig. 3). All of the Family three proteins were identified as plastid-targeted; the program ChloroP (Emanuelsson et al., 1999) predicted transit peptide lengths of 55 amino acids (*Atamy*3), 61 amino acids (*Mdamy*10) and 70 amino acids (*Mdamy*11).

Discussion

Sequencing of apple genes and analysis of sequence databases has yielded a small number of atypical α-amylase sequences. Apple has at least five distinct α-amylase-like genes, grouped into three families; Family two and Family three are both represented by two genes, suggesting that Family one also exists as a pair of genes, one of which has not yet been discovered (the sister gene to *Mdamy*1). This is consistent with the proposed lineage of modern apple, which is thought to have evolved by the fusion of two ancestral genomes, each of which may have contained three α-amylase-like genes, as is the case for *Arabidopsis*.

Perhaps the most striking difference between gene families is the number of introns within each gene and their relative position. Each of the three families has its own characteristic intron structure, with anywhere from two to eleven introns interrupting the coding sequence; Family one and Family three do not share any common intron/exon boundaries, although both share boundaries with Family two. This suggests that, if all three *Arabidopsis* genes have evolved by duplication of a single ancestral gene, then *Atamy*2 is the closest surviving relative in *Arabidopsis*. HUANG et al. (1990), on the basis of the intron structure of cereal α-amylase genes, concluded that intron loss was the dominant pattern of intron evolution in plant α-amylases; however this is not supported by the intron structure of α-amylase genes from apple and *Arabidopsis*. The non-conservation of intron positions between gene families suggests that both intron loss and intron gain has taken place. We expect that all plant α-amylase genes will retain the same family-specific intron structures described here, with gain or loss of one or two introns only.

Very little is known about the expression and regulation of each gene family in either apple or *Arabidopsis*. WEGRZYN et al. (2000) described the upregulation of *Mdamy*8 transcripts in apple fruit held at 0.5 °C for several days. The signal detected in northern analyses is almost undoubtedly a mix of signal from *Mdamy*8 and *Mdamy*9, because the stringency of washing (allowing 14% mismatch) would still detect the *Mdamy*9 transcript, which is 94% identical to *Mdamy*8 in the region of the probe. HILL et al. (1996) showed that cold storage of potato tubers results in the appearance of a new form of amylase within 2–4 days, corresponding with sucrose accumulation in the tuber. This is the same period as shown for upregulation of *Mdamy*8/*Mdamy*9 transcript (WEGRZYN et al., 2000). Unfortunately the potato study does not distinguish between endo- and exo-amylase activity, nor does it look at tuber starch levels during cold storage, but nonetheless it provides an interesting parallel. Potato contains at least one Family two α-amylase, Amy23; expression of this gene was detected in the sprouts of tubers that had been stored at 8 °C for 19 weeks (GAUSING & KREIBERG, 1994).

It is possible that genes from Family two are involved in the general cold acclimation response of plants, which includes an increase in soluble sugars in the cytosol. The α-amylase may have a role in breaking down polysaccharides that could accumulate in the cytosol under these conditions. YANG & STEUP (1990) describe a cytosolic polysaccharide that could be a substrate for cytosolic α-amylases, as well as cytosolic isozymes of other starch-associated enzymes.

A recent microarray analysis of diurnal and circadian regulated genes in *A. thaliana* (SCHAFFER et al., 2001) identified an EST from *Atamy*3 as one of many transcripts displaying a diurnal expression pattern. The *Atamy*3 transcript is upregulated in the afternoon, along with a hexose transporter, and repressed again in the early morning. This pattern corresponds to the diurnal breakdown of starch in chloroplasts and subsequent export of sugars from photosynthetic tissues, which takes place at night, and is consistent with the predicted plastid-localisation of *Atamy*3. The expression pattern for *Mdamy*10 is not known, although its presence in many different tissues has been demonstrated by EST sequencing.

The 5' and 3' untranslated regions of each gene may be important in post-transcriptional regulation, possibly controlling the stability or localisation of the mRNA transcript. The 3' UTR of a rice α-amylase transcript has been shown to mediate mRNA levels in a sugar-dependent manner, by destabilising the transcript when sugar is abundant (CHAN & YU, 1998). Both *Mdamy*8 and *Mdamy*10 have long 3' UTRs that may contain similar mRNA stability elements, and alternate splicing of the 5' UTR intron of *Mdamy*8 may allow different combinations of RNA elements to be expressed.

We believe that all 3 families of α-amylase genes have a different role in plants. Family one α-amylases are involved in the breakdown of extracellular starch, for example in the endosperm of cereal grains, and possibly in the apoplast of virus-infected leaves (HEITZ et al., 1991). Family two α-amylases may have a role in the cytosol,

and could be linked to accumulation of sugars in response to cold temperatures. Family three α-amylases may be the enzymes responsible for degrading all forms of plastid-bound starch, i.e. both diurnal and storage starch. It is possible that other α-amylase gene families will be found in plants; however it appears that the *Arabidopsis* genes described here are sufficient for the fundamental processes of starch metabolism. We are currently attempting to express *Mdamy*8 and *Mdamy*10 in yeast and bacterial systems, to study their enzymatic properties relative to previously-studied α-amylases, particularly from plants. We are also using fusions of *Mdamy*8 and green fluorescent protein (GFP) to study the subcellular localisation of *Mdamy*8 in plant cells, in an attempt to elucidate the function of Family two enzymes.

Acknowledgements

Many thanks to Dr. Richard MacKnight (University of Otago), for his technical advice. We also thank Dr. Andrew Gleave (HortResearch), for providing sequencing facilities, Dr. Ross Crowhurst (HortResearch), manager of the EST sequence database and bioinformatics suite, and Dr. Lesley Beuning (HortResearch), who oversaw tissue collection and library construction for apple ESTs. D.S. is supported by a doctoral scholarship from the New Zealand Agricultural and Marketing Research and Developmental Trust (AGMARDT). Travel assistance was kindly provided by AGMARDT, and the University of Otago Major Research Theme "Functional Genomics, Gene Expression and Proteomics". This work was funded by grants to HortResearch from the New Zealand Foundation for Research, Science and Technology (contracts C06804 and C06X0001).

References

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

Beck, E. & Ziegler, P. 1989. Biosynthesis and degradation of starch in higher plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. **40:** 95–117.

Chan, M.-T., Chao, Y.-C. & Yu, S.-M. 1994. Novel gene expression system for plant cells based on induction of α-amylase promoter by carbohydrate starvation. J. Biol. Chem. **269:** 17635–17641.

Chan, M.-T. & Yu, S.-M. 1998. The 3' untranslated region of a rice α-amylase gene functions as a sugar-dependent mRNA stability determinant. Proc. Natl. Acad. Sci. USA **95:** 6543–6547.

Chen, M.-H., Liu, L.-F., Chen, Y.-R., Wu, H.-K. & Yu, S.-M. 1994. Expression of α-amylases, carbohydrate metabolism, and autophagy in cultured rice cells is coordinately regulated by sugar nutrient. Plant J. **6:** 625–636.

Devereux, J., Haeberli, P. & Smithies, O. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12:** 387–395.

Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. **300:** 1005–1016.

Emanuelsson, O., Nielsen, H. & von Heijne, G. 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci. **8:** 978–984.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:** 783–791.

Felsenstein, J. (1995) PHYLIP (Phylogeny Inference Package).

Gausing, K. & Kreiberg, J. D. (1994) Potato α-amylase gene. European Patent 0 470 145 B1.

Heitz, T., Geoffroy, P., Fritig, B. & Legrand, M. 1991. Two apoplastic α-amylases are induced in tobacco by virus infection. Plant Physiol. **97:** 651–656.

Hill, L. M., Reimholz, R., Schroder, R., Nielsen, T. H. & Stitt, M. 1996. The onset of sucrose accumulation in cold stored potato tubers is caused by an increased rate of sucrose synthesis and coincides with low levels of hexose-phosphates, an activation of sucrose phosphate synthase and the appearance of a new form of amylase. Plant Cell Environment **19:** 1223–1237.

Huang, N., Sutliff, T. D., Litts, J. C. & Rodriguez, R. L. 1990. Classification and characterization of the rice α-amylase multigene family. Plant Mol. Biol. **14:** 655–668.

Jacobsen, J. V., Hanson, A. D. & Chandler, P. C. 1986. Water stress enhances expression of an α-amylase gene in barley leaves. Plant Physiol. **80:** 350–359.

Juge, N., Rodenburg, K. W., Guo, X.-J., Chaix, J.-C. & Svensson, B. 1995. Isozyme hybrids within the protruding third loop domain of the barley α-amylase $(\beta/\alpha)_8$-barrel. Implication for BASI sensitivity and substrate affinity. FEBS Lett. **363:** 299–303.

Kadziola, A., Abe, J., Svensson, B. & Haser, R. 1994. Crystal and molecular structure of barley α-amylase. J. Mol. Biol. **239:** 104–121.

Kadziola, A., Søgaard, M., Svensson, B. & Haser, R. 1998. Molecular structure of a barley α-amylase-inhibitor complex: implications for starch binding and catalysis. J. Mol. Biol. **278:** 205–217.

Langenkamper, G., McHale, R., Gardner, R. C. & MacRae, E. A. 1998. Sucrose-phosphate synthase steady-state mRNA increases in ripening kiwifruit. Plant Mol. Biol. **36:** 857–869.

Lukashin, A. & Borodovsky, M. 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. **26:** 1107–1115.

Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. **10:** 1–6.

Page, R. D. M. 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12:** 357–358.

Pujadas, G. & Palau, J. 2001. Evolution of α-amylases: architectural features and key residues in the stabilization of the $(\beta/\alpha)_8$ scaffold. Mol. Biol. Evol. **18:** 38–54.

Ranjihan, S., Litts, J. C., Foolad, M. R. & Rodriguez, R. L. 1991. Chromosomal localization and genomic organisation of α-amylase genes in rice. Theor. Appl. Genet. **82:** 481–488.

Rodenburg, K. W., Juge, N., Guo, X.-J., Sogaard, M., Chaix, J.-C. & Svensson, B. 1994. Domain B protruding at the third β strand of the α/β barrel in barley α-amylase confers distinct isozyme specific properties. Eur. J. Biochem. **221:** 277–284.

Schaffer, R., Landgraf, J., Accerbi, M., Simon, V., Larson, M. & Wisman, E. 2001. Microarray analysis of diurnal and circadian-regulated genes in Arabidopsis. Plant Cell **13:** 113–123.

Svensson, B. 1994. Protein engineering in the α-amylase family: catalytic mechanism, substrate specificity, and stability. Plant Mol. Biol. **25:** 141–157.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Wegrzyn, T., Reilly, K., Cipriani, G., Murphy, P., Newcomb, R., Gardner, R. & MacRae, E. 2000. A novel α-amylase gene is transiently upregulated during low temperature exposure in apple fruit. Eur. J. Biochem. **267:** 1313–1322.

Yamauchi, D., Takeuchi, H. & Minamikawa, T. 1994. Structure and expression of α-amylase gene from *Vigna mungo*. Plant Cell Physiol. **35:** 705–711.

Yang, Y. & Steup, M. 1990. Polysaccharide fraction from higher plants which strongly interacts with the cytosolic phosphorylase isozyme. I. Isolation and characterization. Plant Physiol. **94:** 960–969.

Received October 4, 2001
Accepted February 12, 2002