EVOLUTION OF MYROSINASE FROM *DRYPETES*

Ntutu Letseka

A dissertation submitted to the School of Molecular and Cell Biology, Faculty of Science, University of the Witwatersrand, in fulfilment of the requirements for the degree of Master of Science.

Johannesburg, 2013

DECLARATION

I declare that this dissertation is my own, unaided work. It is being submitted for the degree of Master of Science to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other institution, and all sources of information have been acknowledged by complete references.

_____

Ntutu Letseka

_20th_ day of _November_, 2013

RESEARCH OUTPUT

Conference Presentations

Letseka, Ntutu and McLellan, Tracy. 'Evolution of Myrosinase from *Drypetes*'. Poster presented at the Bio08 Conference, Grahamstown, 21-25 January 2008.

ABSTRACT

Glucosinolates are a diverse group of molecules found in plants of the order Capparales and the genus *Drypetes*. Hydrolysis of glucosinolates is catalysed by a thioglucosidase, myrosinase. Myrosinase has not only been detected in almost all glucosinolate-containing plants but also in insect and microbial species. Phylogenetic analysis of glucosinolate-containing plants found that all were clustered together with the exception of the outlier genus *Drypetes*. The important question is whether myrosinase in *Drypetes* arose from the same ancestral gene as that in the Capparales, or whether it arose from a different source. Myrosinase-like activity was detected in *D. natalensis*. A candidate molecule for the observed activity was isolated and found to be a 50 kDa heterodimer with subunits of approximately 30 kDa and 20 kDa. This contrasts with Capparales myrosinases which are typically 130-150 kDa homodimers. A 1047 bp partial sequence corresponding to the larger subunit was obtained. Analysis of the nucleic acid and amino acid sequence showed that it was similar to the cupin superfamily of proteins which include the ubiquitous seed storage proteins. Phylogenetic analysis showed that the isolated protein was closely related to seed storage proteins. The results obtained here are suggestive of an independent origin of myrosinase activity in *Drypetes*. With the degree of functional plasticity observed in the cupin superfamily it is proposed that the isolated factor could have acquired myrosinase-like activity. However, purification and further characterisation of the enzyme responsible for the observed activity is still required to confirm the results obtained here.

DEDICATION

This thesis is dedicated to my parents

Matshephe and Moeketsi Letseka

ACKNOWLEDGMENTS

First, I would like to thank the South African National Research Foundation, the University of the Witwatersrand, and the University of South Africa for funding my studies.

I would like to extend my gratitude to my supervisor Prof. Tracy McLellan for her guidance and support. I would like to thank Dr Monde Ntwasa and Professor Rob Veale who have been my supervisors since the retirement of Prof McLellan.

I would like to thank Dr Collet Dandara, in the short time that I knew him for his friendship, understanding and support.

Prof Neil Crouch of the South African National Biodiversity Institute in Durban assisted with the collection of plant material.

I have a lot of good memories of the Population Genetics Lab, my family at WITS. To that I must also add the members of the Fly Lab, Ricardo Antunes, Rodney Hull and Shüné Oliver for sharing the good times and the bad times with me.

Most importantly I would like to thank my family; my wife Helen and my son Elliot for illuminating my life in ways unfathomable, and my brother Tsephe.

Mark Twain once said, **"**There is no sadder sight than a young pessimist.**"** In my pursuit of the most noble of quests, that of knowledge, it is because of the above people that today I can consider myself a converted, happy and well adjusted realist.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

BLAST	Basic local alignment search tool
ESP	Epithiospecifier protein
GOD-Perid	Glucoseoxidase Peroxidase
GSS	Glucosinolate sulphatase
LB	Luria-Bertani
MALDI	Matrix assisted laser desorption/ionisation
MBP	Myrosinase-binding protein
ML	Maximum likelihood
MP	Maximum parsimony
MS	Mass spectrometry
MyAP	Myrosinase-associated protein
NJ	Neighbour joining
NSP	Nitrile-specifier protein
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
Q	Quadrupole
SDS	Sodium dodecyl sulphate
sp	Species
ToF	Time of flight
TFP	Thiocyanate-forming protein
TGG	Thioglucoside glucohydrolases

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Glucosinolates

Glucosinolates are a group of sulphur-containing compounds found in the order Capparales (includes *Arabidopsis* and *Brassica)* as well as the genus *Drypetes* (Rodman et al, 1998). Glucosinolates consist of a *β*-thioglucose group, a sulphonated aldoxime group as well as a variable side chain derived from amino acids (Mithen, 2001). More than 100 glucosinolates have been identified based on the nature of the variable side chain (Chen and Andreasson, 2001). Glucosinolates can be classified into at least 10 chemical classes based on the structure of the variable side chain (Fahey et al., 2001). Side chains include aliphatic, aromatic and heterocyclic groups.

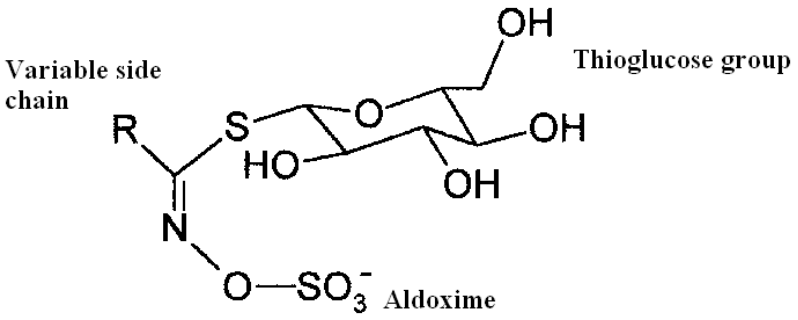


Figure 1.1: General structure of glucosinolates (from Chen and Andreasson, 2001)

Glucosinolates are predominantly inactive molecules that are activated upon hydrolysis by the enzyme myrosinase. Depending on the plant, the pH, the presence of other proteins, and the nature of the glucosinolates themselves, breakdown products can include isothiocyanates, nitriles, epithionitriles and indoles (Wittstock and Halkier, 2002).

1.1.2 Glucosinolate Biosynthesis

The biosynthesis of glucosinolates utilises seven amino acids, namely: leucine, alanine, valine, isoleucine, tyrosine, tryptophan, phenylalanine and methionine (Chen and Andreasson, 2001). This wide pool of precursors gives rise to a large variety of side chain structures, which in turn is responsible for the large variety observed in this group of compounds. Chain elongation of especially methionine, as well as secondary modification of parent glucosinolate side chains, also contributes to the large diversity of glucosinolates observed.

Recent advances have allowed for the partial elucidation of the complex biosynthetic pathway for the generation of glucosinolates. Glucosinolate biosynthesis can be divided into three distinct stages: side chain elongation, generation of the parent glucosinolate, secondary modification of the parent glucosinolate (Wittstock and Halkier, 2002; Mithen, 2001). A brief overview of glucosinolate biosynthesis is given below.

Chain elongation results from the addition of successive methylene groups to methionine and phenylalanine. Two genes have been identified in *Arabidopsis, MAM1* and *MAM-L*, that play a role in elongation of methionine (Kliebenstien et al, 2005). The chain-elongated methionines and the other amino acids are then used to form parent glucosinolates. Parent glucosinolate generation is catalysed by enzymes encoded by two cytochrome P450 gene families, *CYP79* and *CYP83* (Kliebenstein et al, 2005). The amino acid is first converted to an aldoxime by members of the *CYP79* gene family. Five members have been identified, each targeting different amino acids. The aldoxime is then conjugated to a sulphur donor in a reaction catalysed by CYP83. The sulphur donor is enzymatically cleaved by lyases, leaving the sulphur group behind. The last two steps involve the addition of a glucose moiety and a sulphate group by glucosyltransferases and sulphotransferases respectively, to yield the parent glucosinolate (Kliebenstein et al, 2005). Lastly, the parent glucosinolate can undergo extensive side chain modifications. Modification of the side chain occurs in methionine-derived glucosinolates (Wittstock

and Halkier, 2001) and phenylalanine-derived glucosinolates (Chen and Andreasson, 2001). This is mediated by dioxygenases encoded by *AOP* genes in *Arabidopsis* (Kliebenstein et al, 2005).

1.2 Myrosinase

As mentioned previously, myrosinase, a thioglucosidase, is an enzyme capable of catalyzing the hydrolysis of glucosinolates. Since its discovery in 1840, myrosinase has been detected in almost all glucosinolate containing plants (Rask et al., 2000). Myrosinase activity has also been detected in non-plant systems including fungi, bacteria (Bones and Rossiter, 1996) and aphids (Jones et al., 2001; Pontoppidan et al, 2001). A number of studies have aimed at characterizing the properties of myrosinase, its mechanism of action, as well as its evolutionary origins.

1.2.1 General properties of myrosinase

Myrosinase has been extensively studied in the genus *Brassica.* In *B. napus,* myrosinase was shown to exist in homodimeric forms of approximately 130-150 kDa (Rask et al, 2000). Some of these myrosinases were often associated with other proteins in complexes as large as 800 kDa (Lenman et al, 1990). Purified myrosinase has also been shown to be highly glycosylated (James and Rossiter, 1991), with the carbohydrate content consisting mainly of mannose, fucose and N-acetyglucosamine (Rask et al., 2000).

Myrosinases are active across a broad range of temperature and pH values. Myrosinase purified from horseradish was active against its substrate from 23°C up to 50°C (Li and Kushad, 2005). Optimum temperature for activity was between 37°C and 45°C. A myrosinase from *B. napus* expressed in *Saccharomyces cerevisiae* was also shown to function within a broad temperature range with an optimum at 36°C (Chen and Halkier, 1999). Most purified myrosinases have a pH optimum between pH 5.0 and pH 8.0.

Myrosinase from horseradish functioned optimally within this range, with the optimum determined to be pH 5.7 (Li and Kashud, 2005). Myrosinase can also be found in soluble and insoluble forms as has been shown in *Sinapis alba* (Eriksson et al, 2001)

An interesting feature of myrosinase isoenzymes is their activation by ascorbic acid. Myrosinase purified from horseradish was highly activated by ascorbic acid (Li and Kushad, 2005). *Brassica napus* myrosinase is also activated by ascorbic acid (James and Rossiter, 1991). However, very high concentrations of ascorbic inhibited the enzyme to a point where its activity could not be detected.

1.2.2 Myrosinase gene family

Myrosinase in the model plant organism *Arabidopsis thaliana* is encoded by a gene family with three members, named thioglucoside glucohydrolases (*TGGs)* (Bones and Rossiter, 1996). *TGG1* and *TGG2* encode active myrosinases, while *TGG3* was found to be a pseudogene (Zhang et al, 2002). *TGG1* and *TGG2* are approximately 2.6 and 2.7 kb respectively (Xue et al, 1995). Analysis of the TGG3 gene and cDNA sequence revealed frameshift mutations, resulting in transcripts producing truncated proteins (Zhang et al, 2002).

Myrosinase occurs in a number of isoenzymes in *B. napus*, encoded by three gene families, MA, MB, and MC (Rask et al, 2000). In *B. napus,* MA typically encodes a 75kDa myrosinase, MB a 70kDa myrosinase and MC genes encode 65kDa myrosinase (Rask et al., 2000). A number of isoenzymes have been isolated in *Sinapis alba*, belonging to the MA and MB families (Eriksson et al, 2001).

1.2.3 Evolution of myrosinase

Myrosinase is the only known β-*S*-glucosidase and, as such, the evolutionary origin of the enzyme has been of particular interest to researchers. Based on amino acid similarity, myrosinases are grouped with O-glucosidases, in family 1, which includes β-*O*-glucosidases and *β*-galactosidases (Rask et al. 2000). Analysis of the sequences of cynanogenic and non-cyanogenic β-*O*-glucosidases and several myrosinases suggested that myrosinase probably evolved from cyanogenic β-*O*-glucosidases, however the exact relationship between glucosidases is still not clear. Myrosinases, with the exception of one sequence, clustered together. MA, MB and MC myrosinases from *B. napus* formed a distinct clade with each family forming a branch. This cluster was on a separate branch from *Arabidopsis* TGG1 and TGG2 sequences.

1.2.4 Structure and reaction mechanism of myrosinase

The crystal structure of a myrosinase purified from *Sinapis alba* has been elucidated (Burmeister et al, 1997). Myrosinase from *S. alba*, like that of *B. napus* is a homodimer, and is found to be stabilized by zinc ions (Burmeister et al., 1997). Myrosinase shares a common $(\beta/\alpha)_8$ barrel fold with *O*-glucosidases. The monomers interact with one another through a combination of hydrophobic interactions and hydrogen bonds. The structure is stabilised by disulphide bridges, and an extensive number of salt bridges and hydrogen bonds. A number of glycosylation sites were detected confirming earlier reports of myrosinase being a glycoprotein. Myrosinase is extensively glycosylated, with the sugar moieties contributing up to 13kDa per dimer.

While aspects of the mechanism of glucosinolate hydrolysis are not fully understood, the availability of the crystal structure has allowed for great progress in this regard. Analogues of common glucosinolates that act as inhibitors have been used to study the

mechanism of glucosinolate hydrolysis by myrosinase (Iori et al, 1996; Bourderioux et al., 2005). The manner in which ascorbic acid enhances the activity of myrosinase has also been looked at (Burmeister et al, 2000).

Recognition of glucosinolates involves an interaction between the glucose moiety and a glutamic acid residue (Glu-409) residing in the active site of myrosinase (Burmeister et al., 1997). The side chain of the glucosinolate fits into a hydrophobic pocket within the vicinity of the active site. The sulphate group conjugated to the aldoxime is also critical in this process. Indeed, it has been suggested that the sulphate group is the most critical component with regard to binding of the glucosinolate to the myrosinase active site (Iori et al, 1996).

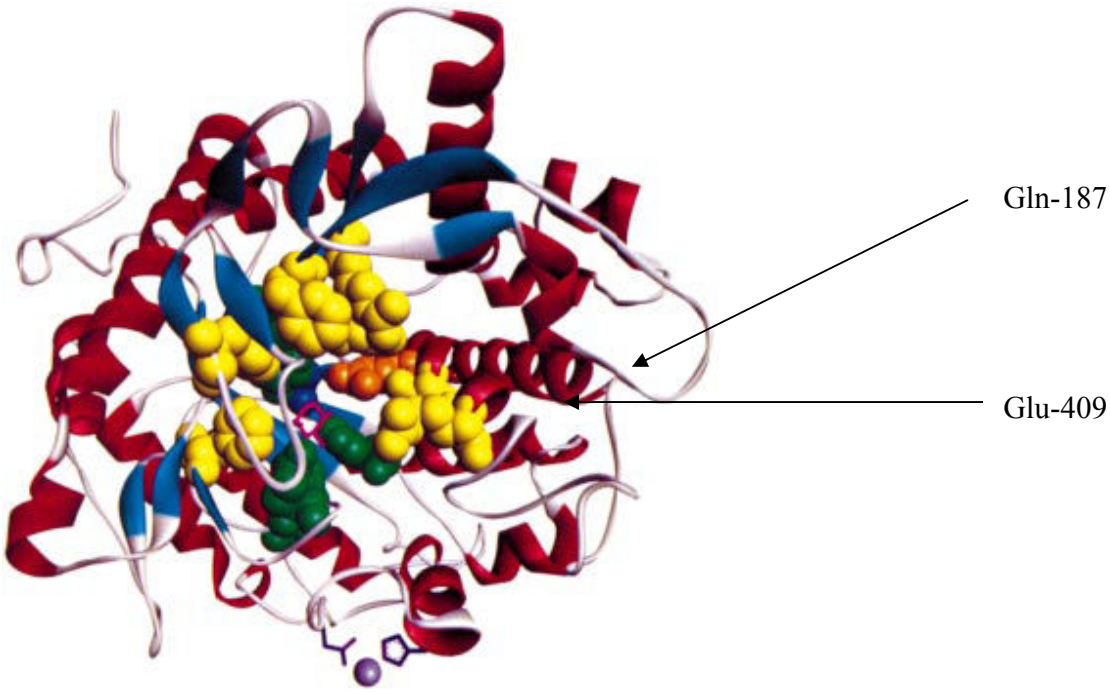


Figure 1.2: Ribbon diagram of a myrosinase monomer from S. alba (from Rask et al., 2000). Residues making up the active site are highlighted (yellow = aglycon recognition, green = glucose ring recognition, cyan = catalytic nucleophile). Ribbons = β-strands, coils = α-helices

The catalytic mechanism is believed to take place in two steps, glycosylation followed by deglycosylation (Burmeister et al., 1997). Glu-409 acts as a nucleophile in the first step, resulting in the formation of the glycosyl-enzyme intermediate as well as the concomitant departure of the aglycon component (side chain and aldoxime) of the glucosinolate. The aglycon, in a non-enzyme mediated process, spontaneously rearranges to form the biologically active by-products of glucosinolate hydrolysis (Bones and Rossiter, 1996). The hydrolysis of the glycosyl-enzyme intermediate is mediated by a nucleophilic water molecule positioned within the vicinity of the active site (Burmeister et al., 1997). The water molecule is brought closer to the substrate by the formation of a hydrogen bond with a glutamine residue (Gln-187) thus facilitating nucleophilic attack. This results in the release of the glycosyl group from the glycosyl-enzyme complex and the regeneration of substrate-free enzyme.

As mentioned previously, myrosinase shows a marked activation in the presence of ascorbic acid. Elucidation of the crystal structure has allowed for the study of this unique feature (Burmeister et al, 2000). Activation of myrosinase by ascorbic acid was believed to be due to co-factor mediated changes in conformation that facilitated increased glucosinolate binding (Bones and Rossiter, 1996). However, parts of the glucosinolates and ascorbic acid bind the same residues in the active site of myrosinase. Burmeister et al. (2000) proposed that ascorbic acid binds only the substrate-enzyme complex following aglycon departure. It acts as a catalytic base, activating the nucleophilic water molecule, which results in enhanced cleavage of the linkage between the glucose moiety and the enzyme.

1.3 Other proteins interacting with myrosinase

Some myrosinase isoenzymes isolated from *B. napus* were found to occur in larger protein complexes. This has intrigued researchers and much work has focused into determining both the nature and function of these proteins. Three classes of proteins have

been found to interact with or function alongside myrosinase, namely myrosinse-binding proteins (MBPs), myrosinase-associated proteins (MyAPs) and specifier proteins.

1.3.1 Myrosinase-binding Proteins (MBPs)

MBPs can broadly be described as proteins that co-precipitate with myrosinase in the presence of anti-myrosinase antibodies (Bones and Rossiter, 1996). The function of these proteins, despite continuous work, remains unknown. What is known about MBPs is that they are not required for the activity of myrosinase. Myrosinase expressed in *S. cerevisiae* was able to function normally without the presence of MBPs (Chen and Halkier, 1999). This myrosinase is normally associated with a number of MBPs in its natural environment. MBPs are, however, the critical component in the formation of myrosinase-MBP complexes (Eriksson et al, 2002). Several MBPs have been identified, with variation occurring in the molecular weight. MBPs from *B. napus* have lectin activity (lectins are proteins that bind carbohydrates). This is critical as lectins have been associated with protection against plant pathogens (Taipalensuu et al, 1997; Chrispeels and Raikhel, 1991). MBPs are also wound- (Taipalensuu et al, 1997) and jasmonate-inducible (Geshi and Brandt, 1998). This suggests a possible role in host-response to pathogens because the accumulation of jasmonate occurs in response to tissue damage or infection by pathogens (Geshi and Brandt, 1998).

1.3.2 Myrosinase-associated proteins (MyAPs)

A second class of proteins that interact with myrosinase are MyAPs. These proteins are unrelated to MBPs (Bones and Rossiter, 1996). Similar to MBPs, these proteins are wound-inducible (Andreasson et al, 1999). A form of the protein has been identified that is induced by methyl jasmonate (Taipalensuu et al, 1997). No specific function has as yet been assigned to MyAPs. However, structural motifs identified have led to the suggestion that MyAP primes glucosinolates for hydrolysis by myrosinase (Andreasson et al, 1999).

Whether MyAPs function only with myrosinase has also been debated since the proteins do not always co-localize to the same cells.

1.3.3 Specifier proteins

The third class of proteins are involved in the specification of the reaction products following hydrolysis of glucosinolates by myrosinase. In the absence of a specifier protein, the aglycone component undergoes a spontaneous structural rearrangement resulting in the formation of isothiocyanates (Bones and Rossiter, 1996). When a specifier protein is present, nitriles, epithionitriles or thiocyanates are preferentially formed at the expense of isothiocyanates (see Figure 3). Two specifier proteins have been identified in plants, epithiospecifier protein (ESP) and thiocyanate-forming protein (TFP).

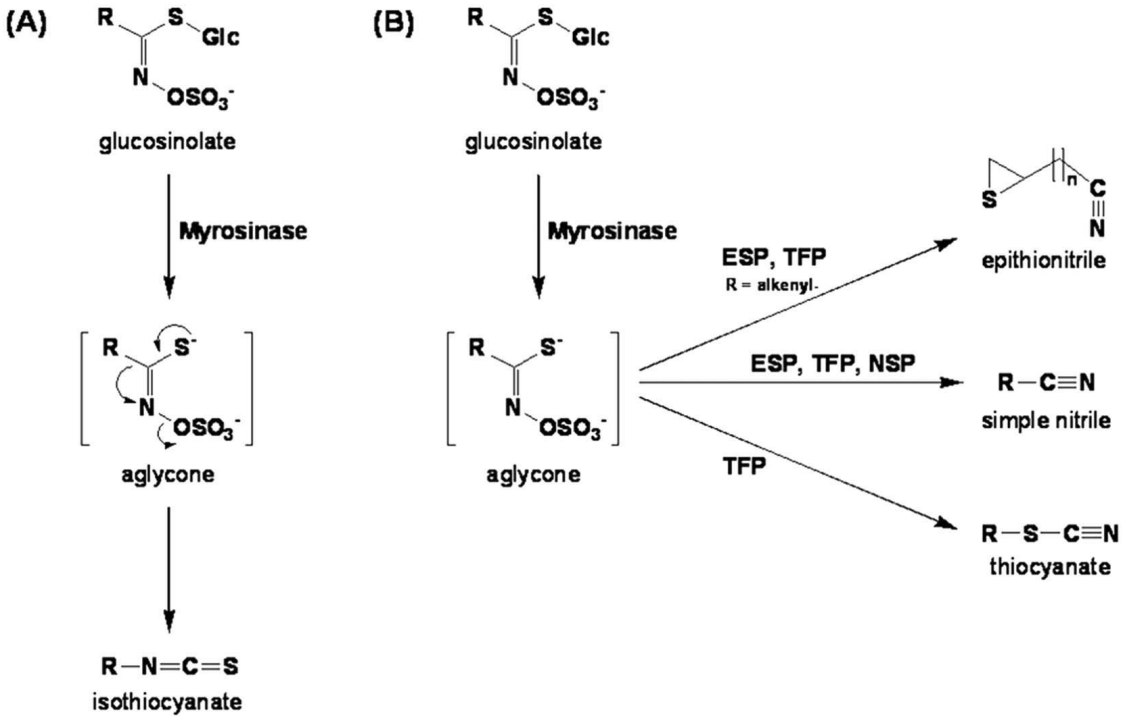


Figure 1.3: Reaction pathway of glucosinolate hydrolysis (A) in the absence of and (B) in the presence of specifier proteins (from Wittstock and Burow, 2007). In the absence of specifier proteins, the aglycone spontaneously rearranges to form isothiocyanates. Specifier proteins redirect the pathway to the formation of simple nitriles, epithionitriles or thiocyanates, depending on the protein present and the R group.

1.3.3.1 Epithiospecifier proteins (ESPs)

ESPs have been identified in a number of plant species which possess glucosinolates and myrosinase (James and Rossiter, 1996). ESP isolated from *B. napus* was found to occur as a 39kDa protein (Bernardi et al., 2000). An additional study also detected a second 35kDa protein which was not thought to be a degradation product of the 39kDa form but instead an isoform (Foo et al., 2000). Epithionitrile formation in the presence of ESP is enhanced by ferrous ions and to a lesser degree the presence of ferric ions. ESP contains a number of protein-protein interaction motifs known as kelch motifs, indicating that

there may possibly be a direct interaction with myrosinase. ESP has also been identified in *Arabidopsis.*

These proteins have no activity towards intact glucosinolates. They were initially believed to affect glucosinolate hydrolysis by binding to and initiating a conformational change in the tertiary structure of myrosinase, thus altering the active site and the substrate-enzyme interaction (Wittstock and Burow, 2007). However, experimental evidence suggests that it is likely that ESPs have a catalytic role to play.

1.3.3.2 Thiocyanate-forming protein (TFP)

TFPs are the most recently identified specifier proteins (Burow et al, 2007). TFP was identified by using *Arabidopsis* ESP to probe a cDNA library generated from *Lepidium sativum*. A complete open reading frame (ORF) of approximately 1020bp encoding a 37 kDa protein was isolated. The DNA sequence and amino acid sequence were 76% and 68% identical to *Arabidopsis* ESP respectively. Thiocyanate formation is limited to a few species including *L. sativum* (Burow et al, 2007) and just three glucosinolates, allylglucosinolate, benzylglucosinolate and 4-methylglucosinolate (Lüthy and Benn, 1977) are believed to have the necessary chemical structure to form thiocyanates. In vitro, TFP redirects glucosinolate hydrolysis to the formation of thiocycanates only when benzylglucosinolate is a substrate. Hydrolysis of allylglucosinolate and 4-methylglusocinolate in the presence of TFP results in the formation of epithionitriles and simple nitriles respectively.

1.4 The glucosinolate-myrosinase system

Together, myrosinase, glucosinolates and other proteins involved form what is referred to as the glucosinolate-myrosinase system. The glucosinolate-myrosinase system is believed to be primarily involved in plant defence (Jones and Rossiter, 1996; Rask et al., 2000).

The products of glucosinolate hydrolysis are toxic to a number of insects, and serve as a deterrent. The system is also believed to be important in plant/pathogen interactions.

1.4.1 The mustard oil bomb

In 1984, Lüthy and Matile proposed the 'mustard oil bomb' hypothesis (Jones and Rossiter, 1996). In this model, myrosinase and glucosinolates are localized to the cytosol and vacuoles respectively. Upon cell damage, glucosinolates are released from the vacuoles and are hydrolysed by myrosinase, releasing toxic products that deter further herbivory. More recent work has disputed the subcellular localisation of myrosinase and glucosinolates as described by Lüthy and Matile, and has led to modifications to this model. Myrosinase is expressed in specific cells called myrosin cells, within specialized bodies called myrosin grains (Kelly et al., 1998; Thangstad et al, 2004). Glucosinolates, on the other hand, are localised to vacuoles in non-specific cells and not in myrosin cells (Kelly et al., 1998). In this new spatial model, it is still necessary for tissue disruption to bring myrosinase and glucosinolates into close proximity.

1.4.2 The glucosinolate-myrosinase system as a defence system

The organisation of the glucosinolate-myrosinase is indicative of a defence tool. Indeed several studies have provided evidence in this regard. Lazzeri et al (2004) showed that products of some glucosinolates were potentially lethal to the root-knot nematode *Meloidogyne incognita*. This was attributed to the production of isothiocyanates.

Recently, *Arabidopsis* double knock-out mutants were generated through T-DNA insertions that were lacking aliphatic glucosinolates (Beekwilder et al, 2008). The genes were involved in regulation of glucosinolate biosynthesis. The effect of this knock-out mutant on insect feeding was tested using the generalist feeder *Mamestra*

brassicae. Insects feeding on mutants grew faster than those growing on wild type plants.

Specifier proteins are also involved with regards to the defensive role of the glucosinolate-myrosinase system. A locus called TASTY was identified in *Arabidopsis* by quantitative trait loci mapping that affected feeding patterns of the insect herbivore *Trichoplusia ni* (Jander et al, 2001). This locus was associated with the production of nitriles in some cultivars, which deterred insect feeding. The locus was found to be very close to the location of the *Arabidopsis* ESP gene (Lambrix et al, 2001).

1.4.3 Other functions of the glucosinolate myrosinase system

Functions of the myrosinase-glucosinolate system are not limited to plant defence against herbivory and pests. Glucosinolates in some cases act as feeding stimulants for specialist feeders (Rask et al, 2000). There has also been great interest in this system due to the anti-carcinogenic properties of some glucosinolate hydrolysis products in the human diet (Bonnesen et al, 2001; Keck and Finley, 2004).

1.5 The myrosinase-glucosinolate system beyond the plant kingdom

While the glucosinolate-myrosinase system has been predominantly found in plants, there have been reports of non-plant myrosinases, most notably in *Brevicoryne brassicae* as well as in fungi and bacteria. Some pests have developed novel strategies to counteract the toxic effect of this system. Some have acquired specifier proteins to divert glucosinolate hydrolysis to less toxic products while others have found ways to inhibit the hydrolysis of glucosinolates by myrosinase. Some of the interactions between the

glucosinolate-myrosinase system and non-plant systems are dealt with in more detail below.

1.5.1 The cabbage aphid, *Brevicoryne brassicae*

1.5.1.1 A novel myrosinase in *Brevicoryne brassicae*

Brevicoryne brassicae was found to possess a myrosinase that was distinct from previously characterized plant myrosinases (Jones et al, 2001). It had already been known that when these insects had been feeding, products of glucosinolate hydrolysis were detectable (MacGibbon and Allison, 1968). This led to the discovery of thioglucosinolate activity in aphids. This however was the first report of a purified myrosinase from any insect species.

Brevicoryne brassicae myrosinase was found to be a homodimer with subunits of 53-54 kDa as estimated by SDS-PAGE and mass spectrometry. Pontoppidan et al (2001) estimated each monomer to be 57-58 kDa, while it was later estimated at 51 kDa (Husebye et al, 2005). Initial peptide sequence analysis showed no relationship to any known myrosinase (Jones et al, 2001). This myrosinase also failed to react with a monoclonal antibody used to detect plant myrosinases (Jones et al, 2001; Pontoppidan et al, 2001). While plant myrosinases respond strongly to ascorbic acid, it was found that the *B. brassicae* myrosinase was inhibited by low concentrations of ascorbic acid, with at most a slight activation as ascorbic acid concentration was increased (Pontoppidan et al, 2001). Another study, however, found its activity was unaffected by ascorbic acid (Husebye et al, 2005). The complete coding sequence was later determined (Jones et al, 2002). This revealed that *B. brassicae* myrosinase was more closely related to animal β-*O*-glucosidases than to plant β-*O*-glucosidases and known myrosinases.

1.5.1.2 Structure of *B. brassicae* myrosinase

The structure of *B. brassicae* myrosinase has been elucidated (Husebye et al, 2005). The structure shows a lot of similarities to other glycosidases. It contains the $(\beta/\alpha)_8$ barrel fold in each monomer that is characteristic of this family. Dimer formation occurs via hydrophobic interactions at the dimer interface. The active site has a binding site for the glucose moiety and an aglycon binding site.

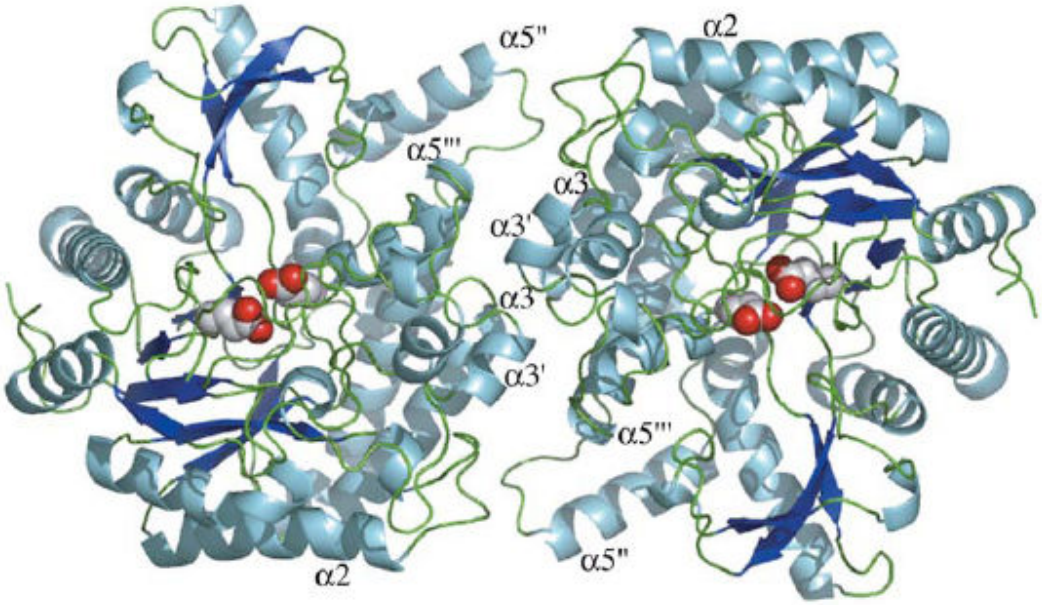


Figure 1.4: Ribbon diagram of the dimeric B. brassicae myrosinase (from Husebye et al, 2005)

Despite a similar structural scaffold, there are some differences to plant myrosinase. Central to plant myrosinases is the substitution of key residues found in β-*O*-glucosidases, with these changes thought to be important for recognition and efficient hydrolysis of glucosinolates. Aphid myrosinase, however, preserves these key residues. The glucose binding is highly conserved but aphid myrosinase maintains a Trp residue that is replaced with Phe in plant myrosinase. The aglycon binding site is similar in size to plant myrosinases but very few residues are conserved.

1.5.1.3 *Brevicoryne brassicae* and the glucosinolate-myrosinase system

Brevicoryne brassicae has evolved a system that is uncannily like that of the glucosinolate-myrosinase system. *B. brassicae* was shown to harbour intact glucosinolates long after feeding (Bridges et al, 2002). The localisation of glucosinolates is unknown but is believed to be separate from myrosinase. Myrosinase is found in muscle tissue while glucosinolates are believed to circulate in the haemolymph. The spatial organisation mirrors that of the glucosinolate-myrosinase system in plants, with the two molecules residing in different organs.

Like in plants, the glucosinolate-myrosinase system in *B. brassicae* is believed to play a role in defence. *B. brassicae* reared on media containing sinigrin produced an isothiocyanate and traces of an alarm pheromone when attacked by ladybirds (Kazana et al, 2007). This effect was specific to sinigrin (Pratt et al, 2008). This was shown to negatively affect the growth of two species of ladybird, inhibiting growth past the first instar stage.

An important aspect to note is that glucosinolates have little or no effect on the growth of *B. brassicae* larvae (Pratt et al, 2008). Thus it is evident that this species has not only adapted to evade its host's defence mechanism, but has acquired a similar mechanism to protect itself from potential predators.

1.5.2 A specifier protein in *Pieris rapae*

A nitrile-specifier protein was identified in the guts of larvae of *P. rapae,* the cabbage white butterfly by Wittstock et al (2004). The authors showed that glucosinolate hydrolysis was diverted from isothiocyanate production to nitrile production by a factor which they called the nitrile-specifier protein (NSP). A 1896 bp ORF encoding a 73 kDa

protein was isolated. Like the plant specifier proteins ESP and TFP, NSP is unable to act on glucosinolates in the absence of myrosinase hydrolysis. The mechanism of action is currently unknown. NSP from the cabbage white butterfly showed no sequence similarity to any known plant proteins, suggesting that this activity had evolved independently as an adaptive measure to counter the toxic effects of glucosinolate hydrolysis. As yet, no equivalent factor has been reported in plants.

1.5.3 Glucosinolate sulfatase in *Plutella xylostella*

The diamond back moth, *Plutella xylostella*, is a crucifer specialist feeder. *P. xylostella* has developed a strategy to avoid the toxic effects of glucosinolate hydrolysis products. An enzyme designated glucosinolate sulphatase (GSS) was discovered in this species (Ratzka et al, 2002). GSS targets glucosinolates directly, removing a sulphate group to form desulpho-glucosinolates (Figure 1.5).

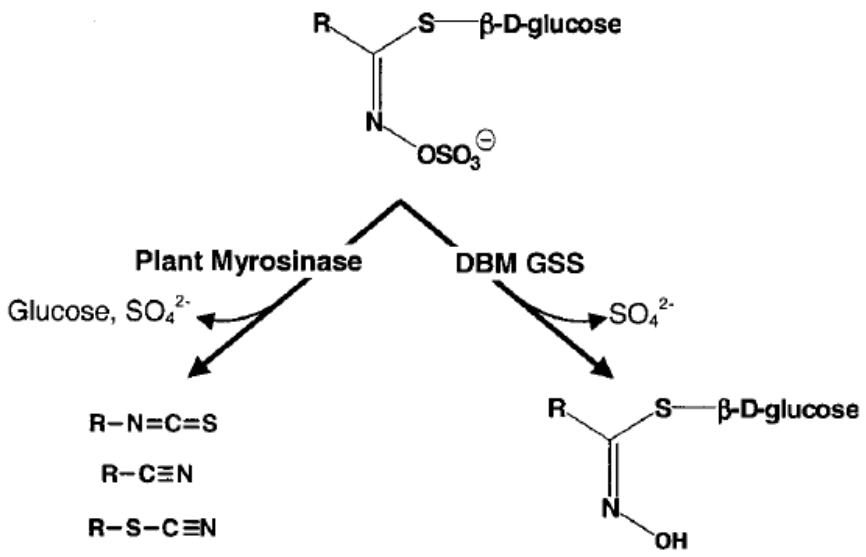


Figure 1.5: Reactions catalysed by myrosinase and diamondback moth GSS (DBM GSS)(from Ratzka et al, 2004). Normally, hydrolysis of glucosinolates by myrosinase results in the formation of isothiocyanates, nitriles and thiocyanates. GSS desulphonates glucosinolates, making them unavailable for the hydrolysis reaction.

GSS is detected in diamondback moth larvae only. Both mRNA transcripts and the active product can be detected in the gut. Myrosinase and glucosinolates come into close proximity in the gut when the larvae feed on plant tissue but GSS is able to desulphonate glucosinolates, which renders them inaccessible to myrosinase, thereby inhibiting the hydrolysis pathway. While it is possible that hydrolysis may take place before GSS acts on glucosinolates, the fact that the diamondback moth successfully feeds on crucifers indicates that this is a good strategy.

1.5.4 Other non-plant myrosinases

Reports of non-plant myrosinase have not been limited to specialist plant feeders. Myrosinase has also been reported in fungi and bacteria (Bones and Rossiter, 1996). The presence of thioglucosidase activity has been demonstrated in *Aspergillus* sp. (Rakariyatham and Sakorn, 2002). The fungus was able to hydrolyze glucosinolates present in brown mustard seed meal, as well as the plant glucosinolate sinigrin. This suggests a possibility of the presence of myrosinase in this fungus, or a myrosinase-like enzyme. The enzyme responsible for the reported thioglucosidase activity has yet to be characterised. Thioglucosidase activity has also been reported in *Sphingobacterium* sp. (Meulenbeld and Hartmans, 2001). This bacterium was capable of hydrolysing several thioglucosides including sinigrin. Unlike plant myrosinases, this thioglucosidase was not activated by ascorbic acid.

1.6 *Drypetes* and the glucosinolate- myrosinase system

The genus *Drypetes* in the family Puntrajivaceae of the order Malpighiales (Rodman et al, 1998, Soltis and Soltis, 2004, Wurdack et al, 2004) occurs in tropical areas throughout the world. A number of species occur in Africa, with at least four species in South Africa (Pooley, 1993). These include *D. reticulata and D. natalensis* (see Figure 6). Several

species are used in some African countries as traditional medicines (Wandji et al, 2000, 2003). Wandji et al. (2000) isolated several compounds from species occurring in Cameroon with known anti-inflammatory properties.

Figure 1.6: *Drypetes natalensis* at the Durban Botanic Garden

1.6.1 The glucosinolate-myrosinase system and *Drypetes*

There has been a report of glucosinolates in *Drypetes* (Kjaer and Friis, 1962), however there have not been any published reports of myrosinase activity. A later study showed the presence of specialist protein-accumulating cells in *Drypetes roxburghii*, but there was no conclusive evidence to suggest that these were myrosin cells (Jørgensen et al,

1977). Later attempts to show the presence of myrosinase activity were unsuccessful (Ettlinger, 1987)

In an analysis of 18S rRNA and *rbc*L sequences all known glucosinolate producing plant species, with the exception of *Drypetes*, were shown to form one major clade (Rodman et al, 1998; Figure 7). The distance between the two clades suggests that this system evolved independently as is most likely the case with non-plant systems. Very little work has been done on the glucosinolate-myrosinase system in *Drypetes* therefore this is still open to investigation.

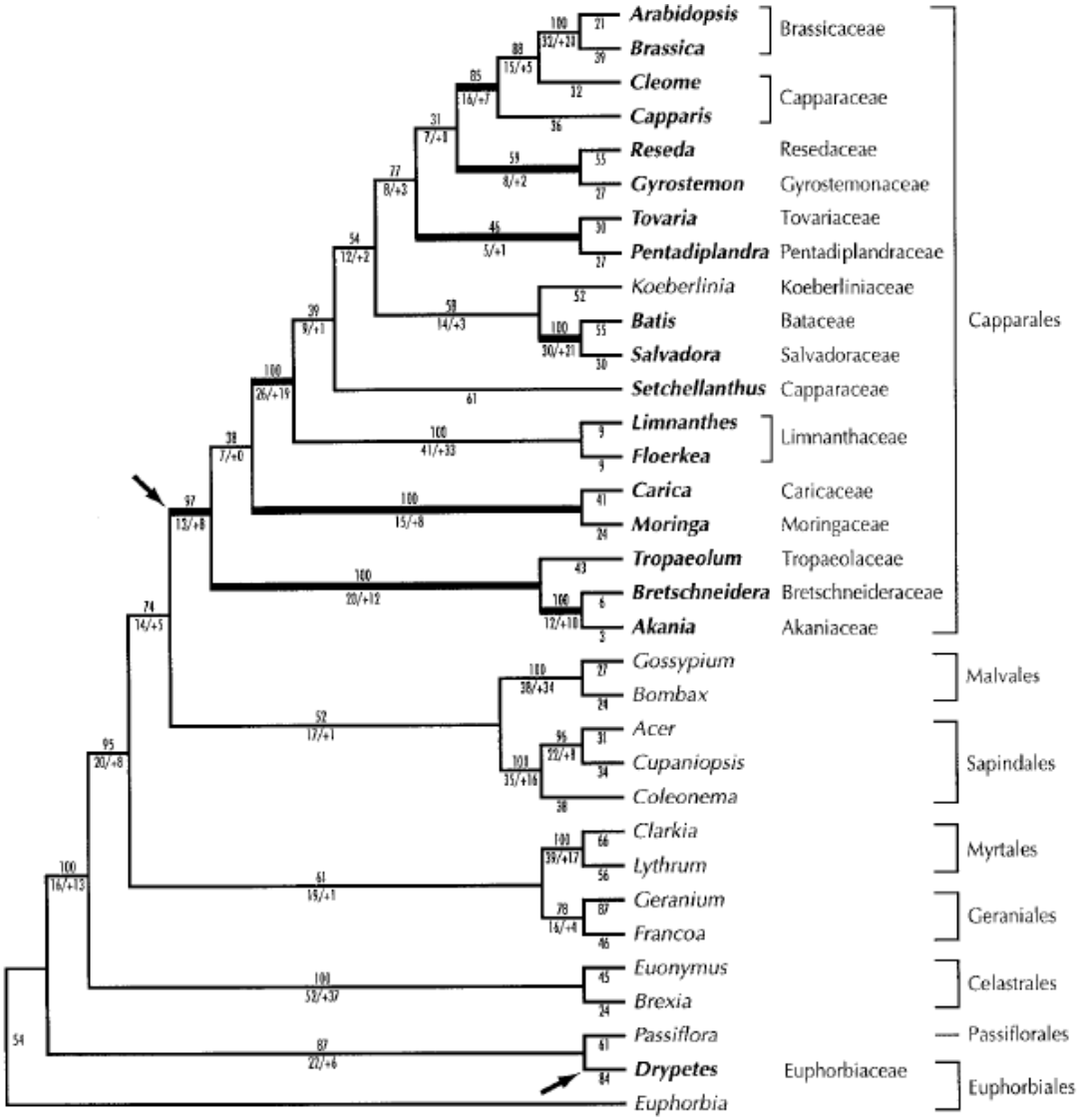


Figure 1.7: Phylogenetic tree indicating the relationship of glucosinolate-producing species (from Rodman et al, 1998).

1.7 Evolution of novel gene function

The origin of novel gene function is a fundamental process in adaptive evolution (Long, 2001). Until the advent of high throughput sequencing methodology and the subsequent archiving of the large volumes of data generated, the underlying mechanisms proposed were merely speculative. Such advances have resulted in a greater understanding and elucidation of the mechanisms involved.

1.7.1 Mechanisms underlying the evolution of novel gene function

The overall scheme followed in the origin of novel genes is that initially a mutation event occurs and providing it is not deleterious it may become fixed over time in a population (Long, 2001). For the most part mutation events that arise will steer a potentially novel gene towards a path of pseudogenisation, a process that eventually leads to the death of the gene (Long et al, 2003). However, in some cases a novel gene function can arise.

There are two schools of thought regarding the process of fixation of novel gene functions (Long et al, 2003). Following the initial mutation event, it has been suggested that purifying selection is relaxed, resulting in the accumulation of random mutations. Later, under selection pressure from, for example, an environmental change these mutations could result in a beneficial function. This is called the 'waiting model'. Conversely, in the 'immediate model', the generation of novel genes can allow for the accumulation of mutations that under the action of positive selection drive the gene towards the acquisition of a new function. Thus while the initial mutation event occurred by chance, subsequent mutation events are selected for and result in the specialisation of gene function.

Several mechanisms implicated in the evolution of novel genes, including duplication events, gene fusion/fission events and exon shuffling are discussed in more detail below.

1.7.1.1 Gene Duplication

The classical method for the generation of novel gene functions is believed to be gene duplication (Long, 2001). Ohno (1973) proposed a model whereby following a duplication event, the duplicated gene no longer undergoes natural selection. Thus over time, the gene accumulates random mutations that will either result in a pseudogene, or in the chance acquisition of a novel function that can become fixed in subsequent generations. This model has been disputed. A model was proposed where gene duplication is preceded by a gene sharing event, thus a single gene having two functions (Hughes, 1994). Following duplication, the sister genes can specialise in one of the functions of the parent gene, a process called subfunctionalisation. Neofunctionalisation occurs in cases where the ancestral gene is not multifunctional (Zhang, 2003). Here, a duplication event could result in a function unrelated to that of the ancestral gene in one of the daughter genes.

There are a number of molecular mechanisms that have been implicated in gene duplication. One such mechanism is retroposition, which results in the random insertion of retrotranscribed mRNA into chromosomal DNA (Zhang, 2003). Retropositioned genes, due to their source material, are usually lacking in non coding sequences or regulatory sequences of their own, and can be characterised by the presence of flanking repeats and/or polyA tracts. In the case of the latter, a requirement for the maintenance of a retropositioned gene in the genome is that it has to be inserted in the proximity of regulatory elements (Long et al, 2003). Retropositioned genes can also form chimerical gene structures and in this way novel functions depending on the environment they insert into. And while a lack of introns

is a signature of retroposition, it must be noted that with time introns can be incorporated in these genes (Long 2001). Likewise, polyA tracts and flanking repeats can be removed.

Unequal crossing over results in the presence of tandem duplicated genes (Long, 2001). Unequal crossing over occurs where there is a misalignment of homologous sequences. While this was thought to be a random process, Metzenberg et al (1991) showed in a study looking at haemoglobin genes in humans that sequence homology is a requirement for unequal crossing over. Segmental duplications are also thought to play a role in the evolution of novel genes (Samonte and Eichler, 2002). Segmental duplications are defined as duplicated regions of DNA 1-200kb with high sequence identity, usually greater than 90%. Unlike duplications generated by unequal crossing over, segmental duplications are not necessarily in tandem. It has been proposed that duplicative transpositions are the most likely causes of segmental duplications.

Lastly, chromosomal and whole genome duplication can be a source for novel genes and is believed to result from non-disjunction of sister chromosomes during mitotic divisions (Zhang, 2003). It is well established that the evolution of the *Arabidopsis* genome has been driven by whole genome duplication events. It was shown that whole genome duplication took place in *Saccharomyces cerevisiae* (Kellis et al, 2004). Comparison to the closely related *Kluyveromyces waltii* showed each region in this species corresponded to two in *S. cerevisiae*. Analysis of gene pairs also showed that one member underwent accelerated evolution.

1.7.1.2 Exon Shuffling

Exon shuffling or recombination, a process that results in the exchange of functional gene segments through non-homologous recombination, was first proposed by Gilbert (1978). Following this proposal, the LDL receptor gene was shown to be a product of

exon shuffling (Sudhof et al, 1985) though the underlying mechanism was as yet unknown. It had been suggested that exon shuffling is an intron mediated process (Gilbert, 1978) due to the presence of transposable elements and repetitive sequences (Kolkman and Stemmer, 2001). Recently, exon shuffling has been attributed to two mechanisms, illegitimate recombination and LINE 1 (L1) element mediated recombination.

Illegitimate recombination takes place between sequences with little or no homology (van Rijk and Bloemendal, 2003). It was initially demonstrated to be a mechanism of exon shuffling by van Rijk et al (1999). They transfected mouse cell lines with a hamster αA-crystallin gene and found that over and above the two normal gene products expected one of the cell lines also stably expressed two larger proteins. Analysis of these products showed that they were a result of an intragenic duplication mediated by illegitimate recombination at CCCAT sequences in intron 3 and exon 2 (which is treated as an intron in 90% of αA-crystallin transcripts). This results in an additional exon 3 in all transcripts resulting in the larger gene product. The CAT sequence has been shown to be a target of topoisomerase I in other cases of illegitimate recombination (Zhu and Schiestl, 1996) suggesting that this enzyme is a critical component of illegitimate recombination mediated exon shuffling.

The L1 element is a retrotransposon that due to a weak transcription termination signal can result in a read through transcript that can allow for genomic elements like exons to be inserted into other parts of the genome (Long, 2001). Moran et al (1999) showed by transfecting human cell lines with L1 that the retrotransposon could affect the shuffling of exons *in vitro*. In practice, for this to result in a novel gene, the L1 transposon would insert in between exon 1 and 2 of a hypothetical gene A. When L1 is transcribed, the weak termination signal allows for a read through transcript that can incorporate all of exon 2. This transcript can then insert into another gene thus incorporating exon 2 in a new position.

1.7.1.3 Gene Fusion/Fission

Gene fusion and fission are molecular events that can contribute to the evolution of novel gene functions. Fusion and fission events have been detected in both prokaryotes (Snel et al, 2000) and in eukaryotes (Thompson et al, 2001; Wang et al, 2004). Gene fusion is thought to occur through the process of readthrough transcription caused by mutations in the transcription termination of the upstream gene (Long et al, 2003). It was shown in humans that a gene *UEV1* on chromosome 20, whose protein product is related to the E2 ubiquitin conjugating enzymes, is expressed as a hybrid with the product of another gene *Kua* (Thompson et al, 2001). The same was not true in worms or flies where *UEV* and *Kua* are distinct genes. The authors hypothesise that due to the presence of a second *UEV* gene on chromosome 8, *UEV1* was a result of a duplication that placed the gene directly downstream of *Kua*. *UEV1* has retained its promoter, however under the direction of the *Kua* promoter, a fused *Kua-UEV1* transcript arises which depending on alternative splicing can result in a truncated transcript producing Kua or a longer transcript producing the fused Kua-UEV1 protein. The product of *UEV1* is localised to the nucleus but when fused with Kua is redirected to the cytoplasm, and the authors suggest that this fusion event increases the scope of substrates that can be targeted for polyubiquitination.

Fission events are rarer than fusion events (Snel et al, 2000). The mechanism underlying gene fission is generally poorly understood. Wang et al (2004) proposed a mechanism they called duplication-degeneration. Using cDNA probes from *D. melanogaster* encoding zinc-finger protein-related sequences, they probed closely related *Drosophila* species and identified a gene family in three species which they called monkey-king (*mkg*), with four signals in *D. mauritiana,* and two in *D. simulans* and *D. sechellia*. All had a common signal on chromosome 3 shared with *D. melanogaster* and additional signals on the X chromosome. Thus they concluded that

signals at chromosome 3 were from the parental genes (designated *mkg-p*). They showed that the additional signals arose from duplication by retropositioning (designated *mkg-r*). Focusing on *D. mauritiana*, they showed that while *mkg-r* and *mkg-r2* maintained a similar transcript structure to the *D. melanogaster* transcript, *mkg-p* and *mkg-r3* deviated significantly. *Mkg-p* showed extensive disruption in upstream coding regions, while *mkg-r3* transcripts were polyadenylated prematurely. Sequence analysis showed that the two shared complementary functions, which were present in a single locus in *D. melanogaster*. The authors thus proposed that following duplication, the two genes were functionally redundant and had undergone gene fission through directed sequence degeneration.

1.7.1.4 Other Mechanisms

De novo origin of whole genes is thought to be rare (Long et al, 2003) but there have been cases where non-coding regions of genes have been converted into coding regions. An example is in the evolution of the *antifreeze glycoprotein* (*AFGP*) gene in Antarctic notothenioid fish (Chen et al, 1997). The authors showed that *AFGP* evolved from an old gene encoding trypsinogen. Sequence alignments showed that the two genes shared 5' and 3' ends with 94% and 96% sequence identity. All of intron 1 of the trypsinogen gene was also present in *AFGP* in two segments. The *AFGP* coding sequence is built on the back bone of a repeating acagcggca (Thr-Ala-Ala) motif, which is present at the splice junction of the trypsinogen gene intron 1 and exon 2. It is believed that the trypsinogen gene's coding sequence was deleted followed by joining of the 5' and 3' ends. Subsequently, the previously non coding acagccggca sequence underwent rounds of replication with the addition of spacer sequences and recruitment of intronic sequences that gave rise to the functional *AFGP* gene.

In prokaryotes, lateral gene transfer has been implicated in the evolution of novel gene function where there is a transfer of non homologous genes (Long et al, 2003). There is a documented case of a product of lateral gene transfer acquiring a novel function in the protozoan *Trichomonas vaginalis* (de Koning et al, 2000). It was shown that *T. vaginalis* possesses a neuraminate lyase gene that was of bacterial origin. It is thought that perhaps acquisition of this gene was a precursor to *T. vaginalis* adopting a parasitic lifestyle, as the gene is found in bacterial parasite, where it is involved in the breakdown of sialic acids, a source of pyruvate.

Also important to note, as in the cases of the gene fusion and fission events described in the preceding section, the mechanisms described are often not observed in isolation (Long et al, 2003). Several mechanisms may be utilised through the evolutionary pathway to ultimately arrive at a functional gene.

1.7.2 Parallel evolution and convergent evolution

Evolution of similar physical or genetic traits in different species takes one of two paths, parallel or convergent evolution. Parallel evolution is defined as the independent evolution of similar traits from a common ancestor (Schluter et al, 2004). At the phenotypic level, this usually refers to the development of a physical trait in closely related lineages (Arendt and Reznick, 2007). At the genetic level, the same building blocks (genes) are used to acquire the same function in independent but usually closely related lineages.

This is in contrast to convergent evolution, which results in shared traits, but not necessarily a shared evolutionary history (Wiens et al, 2004). At the gene level, this implies the sequestering of unrelated genes in different lineages to give the same phenotype. Lactate dehydrogenase (LDH) activity in *T. vaginalis* arose by convergent evolution (Wu et al, 1999). Phylogenetic analysis showed that *T. vaginalis* LDH was

distinct from other known LDHs and evolved through duplication and modification of a malate dehydrogenase gene.

1.8 Problem identification

As mentioned previously, despite the confirmation of the presence of glucosinolates in *Drypetes*, there are as yet no published reports of myrosinase activity. Preliminary results showed the presence of myrosinase activity in three species of *Drypetes* in South Africa (Angus, 2002) as well as the presence of glucosinolates (Rossiter, unpublished results). This is in line with the observation that myrosinase activity was observed in conjunction with the presence of glucosinolates in members of the Capparales (Rask et al, 2000).

Parallel and convergent evolution have always been believed to occur in closely related lineages or distantly related lineages respectively (Arendt and Reznick, 2007). Thus using the principle normally applied, myrosinase activity in *Drypetes,* owing to its distant relation to the Capparales, would most likely have evolved through convergent evolution from a novel ancestral molecule. However, Arendt and Reznick (2007) argued that based on recent evidence this distinction is not necessarily accurate. There have been cases reported that have not followed this generally accepted pathway. Pigment colouration in different populations of the same species of mice was found to have different underlying genetic mechanisms (Hoekstra et al, 2006), and hence a product of convergent evolution. There is also evidence of the opposite, parallel evolution occurring in distantly related lineages. Arendt and Reznick suggest then that it might not be appropriate to base a possible mechanism for evolution of a trait based purely on the phylogenetic distance between species, as they argue is predominantly the case. Leander (2008), however, argued a case for what he called ultimate convergence and proximal convergence (parallel evolution). He argues that in closely related organisms evolution is constrained by homologous networks hence evolution by parallelism, however as genetic distance increases, this constraint is relaxed hence the tendency towards convergence.

As mentioned, myrosinase activity in the Capparales arose by parallel evolution. It is likely that as in the case of aphid myrosinase, myrosinase activity in *Drypetes* is a product of convergent evolution. However, in light of the evidence shown above, the possibility of a parallel origin of myrosinase cannot be discounted. The evolution of the glucosinolate-myrosinase system in plant species poses an interesting question about the evolutionary pathway that was undertaken.

Thus the aims of this project were first, to show the presence of and to characterise myrosinase activity in *Drypetes* and second, to determine its evolutionary origin. This was achieved by isolation of cDNA using a low stringency degenerate primer PCR strategy. Sequencing of the isolated cDNA fragment was followed by phylogenetic analysis using the neighbour-joining, maximum-likelihood and maximum-parsimony methods.

2. MATERIALS AND METHODS

2.1 Sample Collection

Plant material was previously collected in the months of January and February 2001, and December 2002. This includes three species, *Drypetes natalensis, D. arguta* and *D. gerardii.* Additional *D. natalensis* samples were collected on 23 October 2006 and 29 November 2006 at the South African National Biodiversity Institute- Durban. Leaves and fruits dissected into seeds, fleshy pericarp and outer pericarp were preserved in liquid N_2, and then transferred to a freezer at -75°C.

2.2 Protein based methodology

2.2.2 Extraction of total protein from seeds of *D. natalensis*

Crude protein extracts were prepared from stored seeds of *D. natalensis.* Whole seeds were ground to a fine powder with a pre-chilled (two hours at -75°C) pestle and mortar. Approximately 300 mg was transferred to cold (4°C) microfuge tubes and resuspended in 1 ml extraction buffer A (150 mM Tris borate pH 8.9, 0.25% diethyldithiocarbamic acid, 0.5% triton X-100, 5% polyvinylpyrrolidone 40) or extraction buffer B (extraction buffer A, 20% sucrose, bromophenol blue) at 4°C. Samples were spun at 12 500 g for 30 minutes at 4°C. The supernatant was transferred to a fresh microfuge tube and stored at -75°C.

2.2.3 Detection of thioglucosidase activity using the glucose oxidase-peroxidase method

The glucose oxidase-peroxidase (GOD-Perid) method was used as a quick and sensitive method to screen protein extracts for thioglucosidase activity. The success

of the assay is dependent on the presence of thioglucosidase activity in the protein extract. Thioglucosidase activity was determined using the Glucose (GO) assay kit (Sigma). Hydrolysis of glucosinolates by thioglucosidases results in the release of glucose. Glucose oxidase causes the oxidation of glucose into gluconic acid and hydrogen peroxide. The colourless compound o-dianisidine, then reacts with hydrogen peroxide in the presence of peroxidase and is oxidised. Sulphuric acid is added to stabilise this compound resulting in the formation of a pink compound that can be detected spectophotometrically at 540 nm.

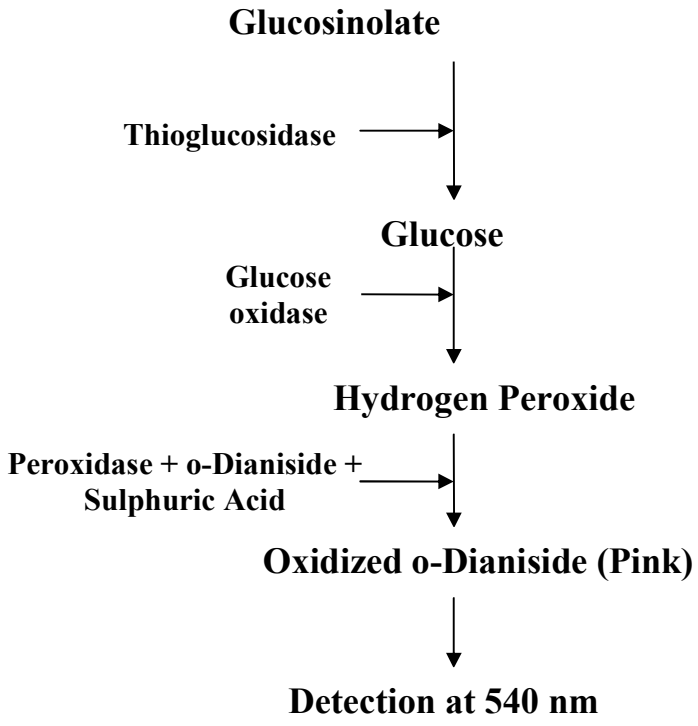


Figure 2.1: An outline of the GOD-Perid method for the detection of glucose liberation from the hydrolysis of glucosinolates by thioglucosidases.

Crude protein extracts prepared in extraction buffer A were dialysed in 200 ml of extraction buffer supplemented with 10 mM EDTA in 1ml aliquots. Extracts were dialysed for 18 hours with buffer changes at six hour intervals. Extracts were diluted 1000 fold prior to assay. Enzyme activity was assayed in 28 mM sodium citrate (pH 5.5) with 4.2 mg/ml sinigrin. Samples were incubated for 30 minutes at 37°C. The

reaction was stopped by incubation for 5 minutes at 95°C. Two volumes of glucose oxidase-peroxidase solution were added. The mixture was incubated for 30 minutes at 37°C. The colour reaction was stopped by adding two volumes of 12 N H_2SO_4. The amount of glucose liberated was determined by comparing the absorbance of the experimental samples to the absorbance of glucose standards at 540 nm.

2.2.4 Detection of thioglucosidase activity using the $BaCl_2$ gel assay

The method used for the detection of thioglucosidase activity using native polyacrylamide gel electrophoresis (PAGE) was developed by MacGibbon and Allison (1970). In summary, the hydrolysis of sinigrin yields sulphate ions which react with $BaCl_2$ to form insoluble $BaSO_4$, which is visible on the gel as a white precipitate. Sites of $BaSO_4$ precipitation are indicative of thioglucosidase activity.

Crude protein extracts prepared in extraction buffer B were used for the $BaCl_2$ gel assay. Seven percent native PAGE gels with a continuous buffer system were used for the detection of thioglucosidase activity. Gels were run in a tris-borate buffer (0.049 M tris, 0.023 M boric acid pH 8.7 (McLellan, 1982)) for 2 hours at 5 V/cm. Gels were stained in 10mM imidazole/HCl pH6.0, 1 mM ascorbic acid, 0.48 mM sinigrin and 60 mM $BaCl_2$ (Bones and Slupphaug, 1989) at 30°C for at least 15 minutes. Gel images were captured using the Bio-Rad and analysed with the Bio-Rad Quantity One software package.

2.2.5 Characterization of myrosinase using sodium dodecyl sulphate (SDS) PAGE

SDS-PAGE was used determine the molecular weight of myrosinase, as well as the size of any subunits. Bands from native gels stained for myrosinase activity were

excised using a sterile surgical blade. The gel fragments were incubated for one hour in SDS loading buffer (63 mM Tris-HCl pH 6.8, 10% glycerol, 2 % SDS; 1% bromophenol blue, with or without 5 % β-mercaptoethanol (Laemmli, 1970)). Gel fragments were run in a discontinuous system with a 4% stacking gel (125 mM Tris-HCl pH 6.8, 0.1% SDS, 0.05% APS, 0.1% (v/v) TEMED) and a 12% separating gel (625 mM Tris-HCl pH 8.8, 1% SDS, 0.05% APS, 0.1% (v/v) TEMED) in electrode buffer (25 mM Tris, 192 mM Glycine, 0.1% SDS) at 10 V/cm for 45 minutes.

Bands were visualised by staining with Coomassie brilliant blue (0.1% Coomasie brilliant blue, 40% ethanol, 10% acetic acid) for 2 hours with gentle agitation. Gels were developed by placing in a destain solution (30% ethanol, 10% acetic acid) overnight. Gels were stored in 5% acetic acid.

Gel images were captured using the Bio-Rad (densitometer). Analysis of the images was done using the Bio-Rad Quantity One desktop application

2.2.6 Sequencing of short peptide fragments

The sensitivity and accuracy of mass spectrometry (MS) has made it a useful tool to identify the primary structure of proteins (Domon and Aebersold, 2006). MS involves the introduction of a simple or complex mixture of molecules into an ionisation source. Two commonly used ionisation sources are matrix assisted laser desorption/ionisation (MALDI) and nanospray ionisation (Nano). The ionised molecules are fed into an analyser where they are separated according to their mass to charge ratio (m/z). Different analysers discriminate between the charged ions differently; time of flight (Tof) analysers determine the m/z according to the time it takes for a molecule to travel along the length of the analyser (Domon and Aebersold, 2006). MS produces limited data allowing protein identification by comparing the protein mass fingerprint generated to a database (Henzel et al, 1993). When two or

more analysers are used in tandem it is possible to generate the primary sequence of the fragments. Charged ions are further fragmented at their amide linkages in a collision chamber in the presence of an inert gas and fed into a second or even third analyser to generate primary sequence data (Hunt et al, 1986).

Two strategies exist to identify proteins using this technology, “bottom up” or “top down”. The first step in “bottom up” involves the chemical digestion of a protein followed by fractionation of the resulting sample using high performance liquid chromatography (HPLC) (Hunt et al, 1988). The short peptide fragments can then be used to generate peptide mass fingerprints (Reid and McLuckey, 2002). If this is inconclusive, MS/MS can be used to determine the amino acid sequence of each peptide fragment. The limitation of this approach is that sequences obtained usually cover just a fraction of the studied protein (Kelleher et al, 1997). The “top down” approach involves no prior digestion of the studied peptide thus allowing for analysis of the entire protein (Reid and McLuckey, 2002; Domon and Aebersold, 2006) but there is a limit to the size of the protein that can be analysed in this manner believed to be an upper limit of 50 kDa.

For the purposes of this project, a “bottom up” approach was used. Bands were excised from SDS-PAGE gels and stored in 5% acetic acid in microfuge tubes. These were sent to the Institut de Biologie Moléculaire et Cellulaire, Strasbourg, France for peptide mass finger print identification and de novo sequencing, through MALDI-Tof MS and Nano-Quadrupole-Tof (Q-Tof) MS/MS respectively.

2.3 Primer Design

Degenerate primers were designed from short peptide sequences. PCR can tolerate primer-template mismatches, with the degree of product yield varying depending on:

a) the nature of the mismatch
b) the position of the mismatch along the length of the primer and
c) the number of mismatches along the length of the primer (Kwok et al, 1990)

Some mismatches (C/C) were found to be more severe than others (A/A) while mismatches at the 3' end of the primer reduced product yield more than mismatches towards the 5' end. As expected, the more mismatches found on a primer, the greater the reduction in product yield.

Designing primers from peptide sequences is a task further complicated by the inherent degeneracy of the genetic code. While codon bias exists for some species this is not always the case. The ability of PCR to tolerate primer-template mismatches allows for the design of degenerate primer pools from peptide sequences. A degenerate primer pool is a mix of primers that incorporates multiple mismatches at ambiguous sites (Kwok et al, 1994). This allows for the incorporation of multiple sequences in a single mix, taking into account the number of codons coding for a given amino acid.

When designing degenerate primers, it is important to take into account the degeneracy of a primer pool, which is the number of unique sequences contained in that pool (Linhart and Shamir, 2005). The greater the degeneracy of the primer, the higher the number of unique sequences present, in effect reducing the concentration of the 'correct' primer. While primers with a very high degeneracy have been used successfully, it is best to keep primer degeneracy as low as possible to obtain maximum product yield and prevent spurious amplification (Linhart and Shamir, 2005)

Several strategies were employed to reduce primer degeneracy. The length of the primer was regulated to ensure that ambiguous positions were kept to a minimum. Longer primers can increase the specificity of the primer by increasing the melting

temperature. In the case of degenerate primers however, this would also increase the likelihood of incorporating multiple mismatches. Where possible, it was attempted to use peptide sequences that had amino acids with low degeneracy's like methionine (1) and tryptophan (2). The use of inosine at the wobble position was also investigated. Inosine is a nucleotide that can form hydrogen bonds with all four bases (Rossolini et al, 1994). Thus incorporating inosine residues into a degenerate primer at select positions can greatly reduce the degeneracy of the primer. However, while inosine is said to be able to bind to *any* other base, the efficiency with which it forms stable base pairs with conventional bases is variable. Watkins and SantaLucia (2005) looked at the stability of I:X pairs as well as the effect of matched adjacent pairs on I:X using nearest-neighbour thermodynamics. They found that the stability of I:X pairs was I:C > I:A > I:T > I:G > I:I. They also showed that the influence of the matched pairs on the 5' and 3' ends of the I:X pair on its stability was as follows G:C > C:G > A:T > T:A. The work of Watkins and SantaLucia (2005) adds another level of complexity to the design of degenerate oligonucleotides as the base composition of both the target sequence and the oligonucleotide is unknown. The primers used in this case had one or two inosines. These were substituted for positions where the degeneracy was 4 fold. By targeting only 4 fold degenerate sites thus limiting the number of inosines used, the aim was to balance reducing the overall degeneracy without compromising the stability of the primer. The least degenerate primer that could be designed for MyrA1 had a degeneracy of 16. The inosine pair, based on sequence data would have been an I:T pair and was flanked on the 5' end by a T:A pair. I:T pairs are not stable, with a positive free energy value, and the T:A pairs at the 5' end of I:X pairs have the most destabilising effect. The authors also showed that I:T pairs were less stable than A:T pairs and suggested than approximating adenosine with inosine is not a good strategy.

It is important to remember that while MS is highly accurate, discrimination between amino acids is based purely on mass. Isoleucine and leucine are structural isomers, with an identical molecular weight, and they are indistinguishable through MS

methods. Similarly lysine and glutamine are of equal mass and molecular weight based methods like MS cannot distinguish between the two. These factors were also taken into account when designing degenerate primers. All possible codons incorporating the variations described above were included when the amino acids responsible were encountered. All primers were purchased from Inqaba Biotec.

2.4 Nucleic acid based methodology

2.4.1 Extraction of total RNA from *Drypetes natalensis* seeds

Total RNA was extracted from seeds of *D. natalensis* using the RedZol (SBS Gentech) reagent, which is based on the guanidine thiocyanate phenol-chloroform RNA extraction method (Chomczynski and Sacchi, 1987). RNA extraction is achieved by phase separation after mixing of an aqueous solution and an acidic solution containing the chaotropic agent guanidine thiocyanate, phenol, and chloroform. Guanidine thiocyanate is a strong protein denaturant (Gordon, 1972) and acts by disrupting bonds that stabilise tertiary structure. At a low pH RNA remains soluble in the aqueous phase, however the solubility of DNA is reduced and thus it remains in the organic phase while denatured proteins are found in the interphase of the biphasic solution. This method thus allows for the preferential extraction of RNA from any contaminating DNA.

Whole seeds were ground to a fine powder with a pre-chilled (two hours at -75°C) pestle and mortar. Approximately 300 mg was transferred to cold (4°C) microfuge tubes and 1.2 ml of room temperature RedZol was added. Samples were spun at 12 500 g for 10 minutes. The supernatant was transferred to a fresh microfuge tube and 240 µl of chloroform was added. The aqueous phase was transferred to a fresh tube. RNA was precipitated by adding 600 µl isopropanol and allowing samples to stand for one hour at room temperature. Samples were spun for 10 minutes at 12 000 g. The supernatant was discarded and samples were washed in 1.2 ml of 70% (v/v) ethanol diluted with DEPC treated water. Samples were spun for five minutes at

440 g. The supernatant was discarded and samples air dried until the ethanol had evaporated. Samples were then resuspended in 50 µl DEPC treated water and quantified by spectrophotometry.

2.4.2 Denaturing agarose gel electrophoresis

2.4.2.1 Sample preparation

Total RNA was transferred into fresh microfuge tubes in 20 µg aliquots. Isopropanol (2.5 volumes) and 3 M sodium acetate (one tenth volume) were added and samples allowed to stand for 10 minutes at room temperature. Samples were spun at 10 000 g and the supernatant was discarded. RNA was resuspended to a final concentration of 1 µg/µl in loading buffer (0.2% bromophenol blue, 6 % glycerol, 6% formaldehyde, 50% (v/v) formamide, 25 mM MOPS, 5mM sodium acetate and 1 mM EDTA).

2.4.2.2 Preparation of denaturing agarose gels

All gel apparatus was sterilised by standing in 20% (v/v) hydrogen peroxide for 1 hour. One percent agarose gels were prepared in MOPS buffer (25 mM MOPS, 5 mM sodium acetate and 1 mM EDTA) diluted in DEPC water. The gel solution was left to stand at 50°C for 1 hour. Formaldehyde was added to a final concentration of 6% (v/v) using a 0.45 µm syringe filter prior to gel casting.

2.4.2.3 Denaturing agarose gel electrophoresis

Samples were run for 4 hours at 4 V/cm. Gels were stained in a freshly prepared ethidium bromide solution (1 ng/ml ethidium bromide and 100 mM ammonium

acetate) with constant shaking at 20 rpm for at least 15 minutes or until bands were visible under UV light. Gels were viewed using the UVP BioDoc-It system.

2.4.3 Extraction of genomic DNA from seeds of *D. natalensis*

Genomic DNA was extracted from leaves of *D. natalensis*. A frozen leaf, approximately 50 ng, was ground to a fine powder with a pestle and mortar stored overnight at -70°C. The powder was transfered to a microfuge tube at 4°C. Genomic DNA was extracted using the Qiagen DNeasy plant mini-kit according to the manufacturer's instruction. The success of the DNA extraction was determined by running the DNA on a 0.8% agarose gel in tris-borate EDTA (TBE) buffer (90 mM tris-borate, 2 mM EDTA pH 8.0). Images were captured with the Bio-Rad GelDoc system using the Quantity One analysis software.

2.4.4 Polymerase Chain Reaction

The polymerase chain reaction (PCR) is a technique that allows for the exponential amplification of a target sequence using primers specific to that sequence (Mullis et al, 1986). The simplicity and robustness of the technique has seen it become one of the most widely used techniques in molecular biology.

2.4.3.1 Reverse Transcription (RT) PCR

Reverse transcription (RT) polymerase chain reaction (PCR) is a modification of the standard PCR for generating complementary DNA (cDNA) from RNA, which can serve as a template for downstream applications. cDNA is generated primarily through the extension of an oligodT primer, which can bind the polyA tail found on

most mRNA species. However, since not all mRNA species are tailed at the 3' end, random generated primers or gene specific primers can be used to facilitate first strand synthesis.

RT-PCR was carried out using the Superscript III Reverse Transcriptase (Invitrogen). Total RNA extracted from seeds of *D. natalensis* was quantified spectrophotometrically and diluted to 2 µg/µl. Into a sterile PCR tube, 5 µg RNA, 0.5 mM of each dNTP and 0.5 µM oligodT or 500 ng random hexamers were added. The mixture was incubated at 65 °C for 5 minutes and immediately transferred to ice. To this mixture, first-strand synthesis buffer (50 mM Tris-HCl pH 8.3, 75 mM KCl, 3 mM $MgCl_2$), 5 µM DTT and 200 U reverse transcriptase in a final reaction volume of 20 µl. For oligodT primers, the mixture was incubated at 50 °C for 1 hour. Samples were quantified and stored at -20°C. Where random hexamers were used, samples were first incubated at 25° for 5 minutes followed by incubation at 50°C for 1 hour. cDNA concentration was quantified by spectrophotometry.

2.4.3.2 PCR

Myrosinase specific cDNA fragments were amplified using gene specific primers designed from short peptide sequences as well as an oligo dT primer. Optimal conditions for each primer pair were determined independently. The parameters looked at were template concentration, primer amount and cycling conditions. To limit the compound effect of multiple parameters each component was tested individually to test whether modification increased the efficiency of the PCR reaction. In general 50 µl reactions were set up with the following components:

Table 2.1: The general PCR cocktail used for amplification of myrosinase fragments

Ingredient	Amount/concentration
cDNA Template	250- 500 ng
Primers	1-4 μM
$MgCl_2$	2mM
dNTP	200 μM of each
Taq polymerase	1.25 U

The cycling conditions were as follows: initial denaturation at 94°C for 2 minutes, 40 cycles of denaturation at 94°C for 30 seconds, primer annealing at 45-50°C for 30 seconds, and extension at 72°C for 30 seconds to 90 seconds, and final extension at 72°C for 5 minutes. All PCR reagents were purchased from Fermenta

The success of PCR was determined by running the reaction products on 1.5% agarose gels in TBE. Images were captured on the Bio-Rad Gel Doc system using the Quantity One (Bio-Rad) analysis software.

2.4.5 Preparation of chemically competent E. coli XL1-Blue cells

Chemically competent *E. coli* XL1-Blue were prepared using the method described by Hanahan (1983). It is generally believed that a combination of low temperatures and the presence of divalent cations is a requirement to induce competency in *E .coli*. Low temperatures reduce the mobility of lipids in the bilayer and the cations serve to shield the negative charges on the phospholipids. Subjecting the cells to heat induces the uptake of DNA into the cells. Whether this is mediated through specific channels is unknown.

E. coli cells were grown in 2 ml cultures overnight in culture tubes in Luria-Bertani (LB) broth (1% tryptone, 0.5% yeast extract, 1% NaCl). The overnight culture was diluted 100 fold in 30-50ml of LB broth. The cell culture was incubated at 37°C with vigorous shaking until an OD reading of between 0.4 and 0.6 was obtained at 600 nm. Cells were decanted into pre chilled 50ml Sorval tubes and left on ice for 15 minutes. Cells were spun for 12 minutes at 2500 r.p.m at 4°C. The supernatant was discarded and cells were resuspended in 1/3 volume of transformation buffer (TFB: 10 mM KMes pH 6.2, 100 mM KCl, 45 mM $MnCl_2$, 10 mM $CaCl_2$, 3mM $HACoCl_3$). Cells were placed on ice for 15 minutes and spun at 2500 r.p.m. for 10 minutes at 4°C. The supernatant was discarded and the pellet resuspended in 1/12.5 volume TFB. DMF was added to a concentration of 3.5% and cells were incubated on ice for 5 minutes. DTT was added to a final concentration of 35 mM and cells were incubated on ice for 10 minutes. A similar volume of DMF was added again to the cell mixture followed by incubation on ice for 5 minutes. Cells were transferred in 200 µl aliquots into pre chilled microfuge tubes. All centrifugation steps were done in a Beckman J2-21 centrifuge with a JA-20 rotor.

2.4.6 Cloning of Myr cDNA fragments into pGEM-T Easy

PCR products were cloned into the TA cloning vector pGEM-T Easy (Promega). TA cloning tales advantage of the fact that any polymerase without 3'-5' proofreading activity like *Taq* polymerase preferentially adds a terminal adenosine to the 3' end of products in the presence of all four nucleotides during the PCR reaction (Clark, 1988). Thus TA cloning vectors are linearised vectors with 5' T overhangs to allow for successful cloning of any PCR product.

Ten microlitre ligation reactions were set up with 50 ng pGEM-T Easy, 25ng-200ng PCR product, 3 Weiss U DNA ligase and ligation buffer (30 mM Tris-HCl (pH 7.8),

10 mM $MgCl_2$, 10 mM DTT, 1mM ATP, 5% polyethylene glycol). Ligations were incubated for at least 16 hours at 4°C.

2.4.7 Transformation of chemically competent *E. coli* cells

2.4.7.1 Transformation of chemically competent *E. coli* JM109 cells

Chemically competent *E. coli* JM109 cells (Promega) were transformed according to the manufacturer's instructions. In summary, 10 ng of recombinant DNA was added to 50 µl of competent cells. For the transformation control 2 ng of pUC18 was added to 100 µl of competent cells. Cells were incubated on ice for 20 minutes, heat shocked for 45 seconds at 42°C and placed on ice for 2 minutes. To each transformation 950 µl LB broth was added (900 µl for the transformation control), followed by incubation for 1 hour at 37°C. Transformations were diluted tenfold in LB both, and 100 µl of diluted and undiluted samples was spread on LB agar (LA) ampicillin plates (LB broth, 1.5% agar, 75 µg /ml ampicillin). Plates were incubated for at least 16 hours prior to clone selection.

2.4.7.2 Transformation of chemically competent *E. coli* XL1-Blue cells

Aliquots of 200 µl of chemically competent *E. coli* XL1-Blue cells were transformed with no more than 25 ng recombinant DNA. To test for transformation efficiency, 2 ng of pUC18 was added to 200 µl of competent cells. Cells were incubated on ice for 30 minutes, heat shocked for 90 seconds at 42°C and placed on ice for 2 minutes. To each transformation 800 µl LB broth was added, followed by incubation for 1 hour at 37°C. Transformations were diluted ten-fold in LB broth, and 100 µl of diluted and undiluted samples was spread on LA ampicillin plates. Plates were incubated for at least 16 hours prior to clone selection.

2.4.8 Small scale preparation of plasmid DNA

Positive clones were selected and sub-cultured on LA ampicillin plates. Plasmid DNA was prepared using the alkaline lysis method (Birnboim and Doly, 1979). This method allows for the separation of circular DNA molecules from linear DNA and protein contaminants. Cells in Tris-EDTA buffer are lysed in a solution containing SDS and NaOH. The increase in pH also results in denaturation of large macromolecules like linear chromosomal DNA. However, circular DNA retains its structural integrity. The solution is neutralised by adding an acidic acetate solution. Linear DNA renatures but due to its high molecular weight, it aggregates and precipitates out along with proteins. The contaminants are removed by centrifugation at high speeds. Contaminating RNA can be removed by treating with RNase A.

Escherichia coli cells were grown in 2 ml cultures overnight in sterile LB broth in culture tubes. Cells were transferred to 2.2 ml collection tubes and spun at 12 500 g for 30 seconds. The supernatant was discarded and cells were resuspended in 100 µl solution 1 (50 mM glucose, 25 mM Tris-HCl pH 8.0, 10 mM EDTA). Immediately, 200 µl solution 2 (0.1 M NaOH, 1% SDS) was added and the suspension mixed by inverting the tubes. Samples were incubated on ice for at least 5 minutes until the solution was clear. One hundred and fifty microlitres of solution 3 (3 M sodium acetate) was added and the samples were mixed by inverting. After incubating on ice for 5 minutes, samples were spun at 12 500 g for 10 minutes and the supernatant transferred to a fresh microfuge tube.

To precipitate any remaining protein contaminants, an equal volume of phenol:choloroform:isoamyl alcohol (24:24:1) was added and samples were spun at 12 500 g for 1 min. The aqueous phase was transferred to a fresh microfuge tube. An equal volume of chloroform was added and samples spun at 12 500g for 1 min. The aqueous phase was transferred to a fresh microfuge tube.

To precipitate plasmid DNA an equal volume of isopropanol was added. Samples were allowed to stand at room temperature for 5 minutes and were then spun at 12 500 g for 10 minutes. The supernatant was discarded, then 1 ml 70% ethanol was added. Samples were spun at 12 500g for 2 min. The supernatant was discarded and samples were air dried to remove the residual alcohol. Plasmid DNA was resuspended in 50 µl Tris- EDTA buffer (10 mM Tris-HCl pH 7.4, 1 mM EDTA, 2ng/ml RNase A) and incubated at 65°C for 20 minutes. The purified plasmid DNA was stored at -20°C.

2.4.9 Screening of purified plasmid DNA for the presence of an insert

Purified plasmid DNA was screened for the presence of an insert using restriction analysis. The pGEM-T Easy vector contains two sites for the restriction endonuclease *Eco*RI either side of the PCR product cloning site allowing for the release of inserts with a single digest. No more than 1 µg of plasmid DNA was incubated with 5 U *Eco*RI and *Eco*RI buffer (50 mM NaCl, 100 mM Tris-HCl, 10 mM $MgCl_2$, 0.025% Triton X-100 (pH 7.5)). Digestions were incubated for 16 hours at 37° and then separated on 1.5% agarose gels in 1X TBE. Images were captured with the Bio-Rad GelDoc system and Quantity One analysis software. *Eco*RI was purchased from New England Biolabs.

2.4.10 Sequencing

Purified plasmid DNA with an insert of interest was sent to Inqaba Biotec for sequencing. Samples were sequenced using the universal SP6 and T7 as forward and reverse primers. For large inserts in excess of 1 Kb, an internal sequencing primer was designed to obtain the full sequence of the clone.

2.5 Bioinformatics

2.5.1 Sequence analysis

Sequence analysis was done using the desktop application Sequencher ® (Gene Codes Corp). Sequences obtained were aligned with two short (approximately 100bp) pGEM-T Easy sequences, either side of the cloning site, to identify vector sequences, which were trimmed. The remaining nucleotide sequence was translated in all three frames. To confirm that the sequence was of the correct origin, the peptide sequences corresponding to the primer pair used was compared to the translated sequences.

2.5.2 Protein sequence analysis

Three online tools were used to search for conserved domains in the translated amino acid sequence: (a) Pfam protein family database (Bateman et al, 2004); (b) NCBI Conserved Domain Database (Marchler-Bauer et al, 2007) and (c) ScanProsite (Sigrist et al, 2010). ProtParam (Gasteiger et al, 2005) was used to estimate the molecular weight of the translated amino acid sequence. The web locations for the tools mentioned above are listed in Table 2.2.

Table 2.2: Web locations of tools used for used for protein sequence analysis

Database	URL
Pfam protein family databse	http://pfam.sanger.ac.uk/
NCBI conserved domain database	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
ScanProsite	http://www.expasy.ch/tools/scanprosite/
ProtParam	http://ca.expasy.org/tools/protparam.html

2.5.3 Sequence identification

The Basic Local Alignment Search Tool (BLAST) (Altschul et al, 1990) from the National Centre for Biological Information (NCBI) was used to identify any matches with the sequence obtained. Two algorithms were used to identify related sequences. Blastn is an algorithm that looks for somewhat similar sequence matches in the nucleic acid database while Blastp searches for similar protein sequence matches.

2.5.4 Phylogenetic analysis

2.5.4.1 Multiple sequence alignments

Sequences with a low E value (less than 0.01) were downloaded from GenBank. Multiple sequence alignments were completed using the desktop application ClustalX 2.0 (Larkin et al, 2007). This was done for nucleic acid sequences and amino acid sequences.

2.5.4.2 Neighbour-joining (NJ) analysis

The NJ method estimates the evolutionary distance between aligned sequences as a function of the number of changes or mutations between sequences, with the data presented in a distance matrix (Holder and Lewis, 2003). The aim of the method is to arrive at a final tree which can be explained using the principle of minimum evolution (Saitou and Nei, 1989). The principle of minimum evolution states that the tree with the smallest sum of branch lengths is the correct tree. Neighbour-joining analysis is not a computationally intensive method and it can underestimate evolutionary distance between highly divergent sequences (Holder and Lewis, 2003).

Phylogenetic trees inferred using the NJ method were constructed using Clustal X 2.0. Trees were graphically displayed using NJplot (Perrière and Gouy, 1996).

2.5.4.3 Maximum parsimony (MP) and maximum likelihood (ML) analysis

MP and ML are discrete methods that infer phylogenies by doing a site-by-site comparison as opposed to computing distances between sequences (Page and Holmes, 1998).

MP discriminates between phylogenetic trees by selecting the tree which requires the minimum amount of changes to explain a given data set (Felsenstein, 2004; Steel and Penny, 2000). Thus MP does not necessarily require prior knowledge of any evolutionary model that could have produced a given data set. MP is a relatively simple method for inferring phylogenies.

The most parsimonious tree will not always be correct since more complex evolutionary pathways do occur that cannot be explained by the principles of parsimony (Stewart, 1993). In that case, a method that incorporates evolutionary models would be preferred. ML is a method that judges a tree on how well it predicts a given data set (Holder and Lewis, 2003). For each possible tree that can be constructed, the probability of it matching the data is determined (the likelihood score), and the tree with the highest probability is selected. Unlike MP and NJ analysis, this method allows for the incorporation of models of evolution into the estimate of the tree. This does however make this method the most computationally taxing for very large data sets.

MP and ML analysis was done using the Phylogenetics Inference Package (PHYLIP) 3.68 (Felsenstein, 2005). PHYLIP is a package consisting of a number of individual

programs for the inference of phylogenies. Bootstrap analysis with 100 replicates was first done using the program seqboot. MP analysis was performed using the program Dnapars and Protpars for nucleic acid sequences and amino acid sequences respectively. ML analysis was done using Dnaml and Proml. Following analysis, a consensus tree was computed using the program consense. Trees were graphically displayed using NJplot (Perrière and Gouy, 1996).

3. RESULTS

3.1 Detection, characterisation and identification a thioglucosidase in *D. natalensis*

3.1.1 Screening of tissues for thioglucosidase activity

Seeds collected at different time points were screened for thioglucosidase activity. While preliminary results have indicated the presence of thioglucosidase activity in this species (Angus, 2002), it is not known at which point during development expression takes place. Thus seeds from fruits less than 2cm in diameter were collected in October 2006 and more mature seeds up to 4 cm in diameter collected in November 2006. Seeds collected previously were also screened for thioglucosidase activity. Leaves from male and female trees from the 23 October 2006 collection were also screened for enzyme activity.

Enzyme activity was detected in seeds four centimetres in diameter from 29 November 2006 collections. However, activity could not be detected in smaller, presumably less mature seeds from this time. Enzyme activity was detected in seeds from fruits 3-4 cm in diameter collected on 8 December 2002. No enzyme activity was detected in seeds dissected from early fruits. There was also no detectable enzyme activity from either leaves from male and female trees. The majority of tissues from earlier collections yielded no enzyme activity with the exception of green fruits more that 2 cm in diameter

Therefore seeds from large fruits (samples 523 and 544) were used for further work. These results are indicated in Table 2.1.

Table 3.1: Summary of the result of screening various tissues of *D. natalensis* for thioglucosidase. Tissue collected at different time points, thus at different levels of maturity, were screened for thioglucosidase activity.

Collection Number	Collection Date	Location	Description	Activity (y/n)
523	8 Dec 2002	Sodwana Bay	Seeds from green fruits, 3-4 cm diameter	y
524	8 Dec 2002	Sodwana Bay	Seeds from slightly orange fruits, 4 cm diameter	n
525	8 Dec 2002	Sodwana Bay	Seeds from green fruits, 2 cm diameter	n
526	8 Dec 2002	Sodwana Bay	Fruit (orange)	n
539	23 Oct 2006	Botanic Gardens, Durban	Young leaves from a female plant	n
541	23 Oct 2006	Botanic Gardens, Durban	Young leaves from a male plany	n
543	23 Oct 2006	Botanic Gardens, Durban	Seeds from 1-2cm green fruit, from tree next to glasshouse near herbarium	n
544	29 November 2006	Botanic Gardens, Durban	Seeds from 4 cm green fruit, from tree next to glasshouse near herbarium	y

3.1.2 Detection of thioglucosidase activity in *D. natalensis*

Thioglucosidase activity was detected using a gel assay specific for this activity. Activity was detected in extracts prepared from sample 544. A white Ba_2S0_4 precipitate was visible within 10 minutes at 30° (see Figure 3.1). Despite the GOD-Perid assay indicating the presence of enzyme activity in extracts prepared from sample 523, activity could not be detected using the gel assay, even when left for up to 16 hours in the $BaCl_2$ staining solution at 30°C or 37°C. For this reason all further work was done using sample 544.

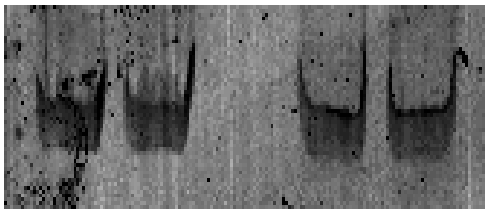


Figure 3.1: $BaCl_2$ Assay for the detection of myrosinase activity after nondenaturing gel electrophoresis. Hydrolysis of the naturally occurring glucosinolate sinigrin, results in the liberation of SO_4^{2-} ions which displace Cl^- to form a white $BaSO_4$ precipitate.

3.1.3 Characterisation of the enzyme responsible for thioglucosidase activity in *D. natalensis*

Bands from native gels were excised and loaded onto SDS-PAGE gels. This was done to determine the molecular weight of the protein. Fragments were incubated in loading buffer with or without the denaturant β-mercaptoethanol, to determine the size of any subunits. In non-denatured samples, a single band was detected, corresponding to a molecular weight of approximately 50 kDa. Upon denaturation,

two bands were detected of approximately 30 kDA and 20 kDa, designated Myr A and Myr B respectively (Figure 3.2)

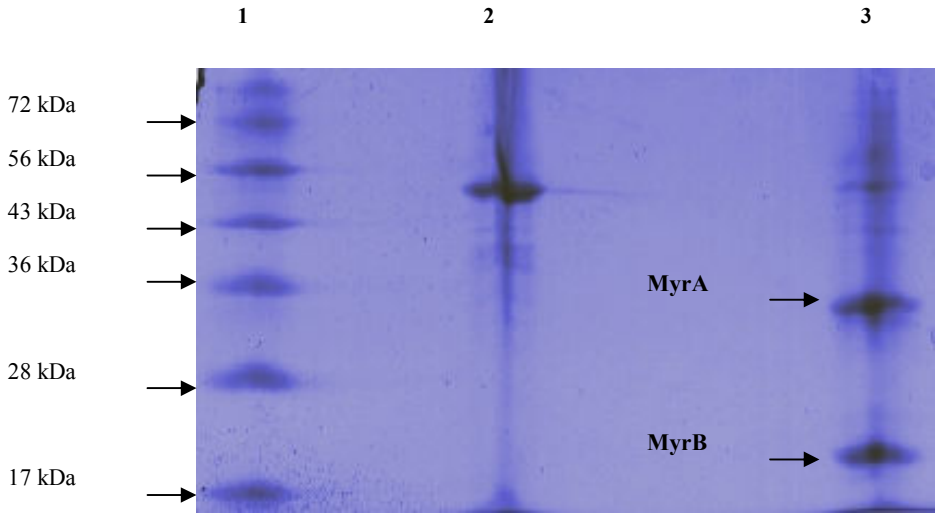


Figure 3.2: SDS-PAGE Analysis of Myrosinase. Lane 1 = Marker, Lane 2 = Untreated gel extract, Lane = β-Mercaptoethanol treated gel extract. Native myrosinase migrated as a 50 kDa fragment. Reduced myrosinase migrated as two fragments of approximately 30 kDa (MyrA) and 20 kDa (MyrB).

3.1.4 Sequencing of myrosinase fragments

Myr A and Myr B were excised from SDS-PAGE and sent to the Institut de Biologie Moléculaire et Cellulaire, Strasbourg, France, for identification and de novo peptide sequencing, through MALDI-Tof MS and Nano-Quadrupole-Tof MS/MS. The peptide mass fingerprint generated using MALDI-Tof MS produced no matches to any known proteins. The fragments were then subjected to Nano-Quadrupole-Tof MS/MS to generate primary sequence data. Seven short peptide sequences and three short peptide sequences were obtained for Myr A and Myr B respectively, ranging from six amino acids to 14 amino acids in length.

Table 3.2: Short peptide sequences obtained following peptide sequencing by tandem mass spectrometry of MyrA and MyrB

Subunit Name	Short Peptide Name	Short Peptide Sequence
MyrA	MyrA1	QGQLDNNPR
	MyrA2	QVVGPTPR
	MyA3	NNLFYGLNAK
	MyrA4	QLQSLLAR
	MyrA5	LLAEALNLDENLAR
	MyrA6	VSLLNSQNLVPLR
	MyrA7	DLEVLANAY
MyrB	MyrB1	DNAALA
	MyrB2	REETLC
	MyrB3	ADVFNPDQR

A BLAST search showed that short peptide sequences from Myr A were similar to sequences found in a legumin-like protein from *Ricinus communis*. Sequences from Myr B were similar to an unnamed 11S storage protein precursor.

3.2 Amplification, cloning and sequencing of cDNA corresponding to Myr A and Myr B

3.2.1 Primer Design (a)

Degenerate primers were designed from short peptide sequences (Table 3.2). The initial strategy employed was to use 17mer oligonucleotide with inosine substituted at positions with 4 or more possible nucleotides. This had the benefit of reducing the degeneracy of the primer 4-fold for each inosine used. For example MyrB1 contains a

single inosine and the resultant primer is 32-fold degenerate. Had inosine not been used the primer would have been 128 fold degenerate. Three peptides were selected for Myr A and B that had the lowest degeneracies.

Table 3.3: Primer sequences designed for RT-PCR amplification of *Drypetes* myrosinase mRNA.

Subunit Name	Primer Name	Primer Sequence	Degeneracy
MyrA	MyrA1	CARCTIGAYAAYAAYCC	16
	MyrA2	GARGTNCTIGCNAAYGC	64
	MyrA3	TCICARAAYCTYGTNCC	32
MyrB	MyrB1	GAYAAYGCIGCNCTYGC	32
	MyrB2	AGRGARGARACICTNTG	32
	MyrB3	GCIGAYGTITTYAAYCC	8

I=Inosine, Y= C or T, R=A or G, N=A, C, G or T

3.2.2 Amplification of Myr fragments using inosine primers

Inosine primers were used in conjunction with an 18mer oligodT primer. No amplification was observed even at an annealing temperature as low as 35°C with 40 cycles.

3.2.3 Primer Design (b)

Following the lack of amplification using the first set of primers, more primers were designed (Table 3.3). Inosine was substituted for the naturally occurring nucleotides. To increase the likelihood of success each peptide sequence was used to design to design two primers, one in the forward direction and its reverse compliment. This was based on the fact that sequential order of each peptide fragment in each subunit was unknown. In effect there are 24 possible orientations of the peptide fragments

used to design primers from MyrA. These can be divided into four reaction sets depending on which fragment occurs first in the sequence. There are a possible six for MyrB, grouped into three reaction sets. Using the primers in combination negated the lack of prior knowledge relating to the order of the peptide fragments. The primers were used in combination as indicated in table 3.4 (a) and (b).

Table 3.4: Additional primer sequences designed for PCR amplification of *Drypetes* myrosinase mRNA.

Subunit	Peptide Sequence	Forward Primer	Reverse Primer	Degeneracy
MyrA	QLDNNPR	MyrA1f: MARYTNGAYAAYAAYCC	MyrA1r: GGRTTRTTRTCNARYTK	128
	NNLFYGL	MyrA3f: AAYAAYYTNTTYTAYGGNYT	MyrA3r: ARNCCRTARAANARRTTRTT	256
	ELNLDE	MyrA5f: GARYTNAAYYTNGAYGA	MyrA5r: TCRTCNARRTTNARYTC	512
	EVLANA	MyrA7f: GARGTNYTNGCNAAYGC	MyrA7r: GCRTTNGCNARNACYTC	512
MyrB	DNAALA	MyrB1f: GAYAAYGCNGCNYTNGC	MyrB1r: GCNARNGCNGCRTTRTC	512
	REETLC	MyrB1f: GAYAAGCNGCNYTNGC	MyrB2f: CANARNGTYTCYTCNCK	128
	FNPDQR	MyrB3f: TTYAAYCCNGAYCARMG	MyrB3r: CTYTGRTCNGGRTTRA	128

Y= C or T, R=A or G, M=A or C, K=G or T, N=A, C, G or T

Table 3.5: Workplan for the amplification of (a) Myr A and (b) Myr B
(a)

	MyrA1r	**MyrA3r**	**MyrA5r**	**MyrA7r**
MyrA1f	-	Y	Y	Y
MyrA3f	Y	-	Y	Y
MyrA5f	Y	Y	-	Y
MyrA7f	Y	Y	Y	-

(b)

	MyrB1r	**MyrB2r**	**MyrB3r**
MyrB1f	-	Y	Y
MyrB2f	Y	-	Y
MyrB3f	Y	Y	-

- = Not tested, Y = Tested

3.2.4 Amplification of Myr B fragments

The use of a new set of primers failed to yield any PCR product. No amplification was obtained when the three forward when used with a reverse olidodT primer. Attempts to amplify Myr B fragments from genomic DNA were unsuccessful. Amplification of Myr B was thus abandoned.

3.2.5 Amplification of Myr A fragments

Four peptide sequences were used to design primers for Myr A as indicated in Table 3.3. This resulted in a possible 12 reactions that could be attempted. Out of all these reactions fragments were obtained for three combinations of primers.

Amplification of Myr A fragments from genomic DNA was unsuccessful. Multiple products were obtained with some of the primer pairs while others yielded no

product. Multiple products could not be duplicated with repeated attempts. Amplification of Myr A from genomic DNA was thus abandoned.

Amplification with MyrAlf and MyrA3r from cDNA yielded a fragment of approximately 170 bp (Figure 3.3). Amplification with MyrA1f and MyrA7r yielded a fragment of approximately 380 bp (Figure 3.3). The optimum reaction conditions for these two reactions were: initial denaturation at 94°C for 2 min, 40 cycles of denaturation at 94°C for 30s , annealing at 50°C for 1 min, extension at 72°C for 1 min, and final denaturation at 72°C for 1 min.

Amplification with MyrA1f and MyrA5r from cDNA yielded a fragment of approximately 470 bp (Figure 3.3). The optimum reaction conditions for this reaction was as follows: initial denaturation at 94°C for 2 min, 40°C cycles of denaturation at 94°C for 30s , annealing at 45°C for 1 min, extension at 72°C for 1 min, and final denaturation at 72°C for 1 min. These results are indicated in figure 3.3.

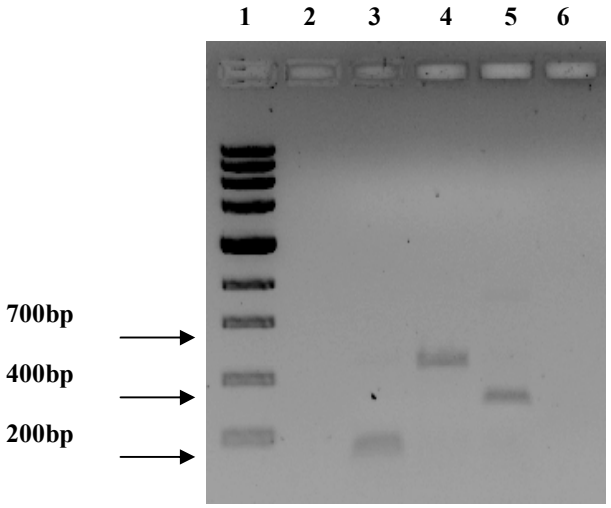


Figure 3.3: Amplified Myr cDNA fragments. Fragments were amplified using a combination of degenerate primers designed from peptide sequences obtained following sequencing. PCR products were electrophoresed on 1.5% agarose gels. Lane 1 = Molecular weight marker, Lane 3 = MyrA1f/MyrA3r, Lane 4 = MyrA1f/MyrA5r, Lane 5 = MyrA1f/MyrA7r

Since all three fragments were amplified with MyrA1f as a common forward primer, all further work was done using the largest fragment, amplified with MyrA1f/MyrA5r, hereafter designated MyrA1.5. This was because fragments obtained using MyrA3r and MyrA7r should be contained in the MyrA1.5 fragment based on the size of fragments.

3.2.6 Cloning and sequencing of MyrA1.5

MyrA1.5 was cloned into pGEM-T Easy. Restriction analysis was performed by digesting the purified plasmid DNA with *EcoR*I to confirm that the correct fragment had been cloned (Figure 3.4).

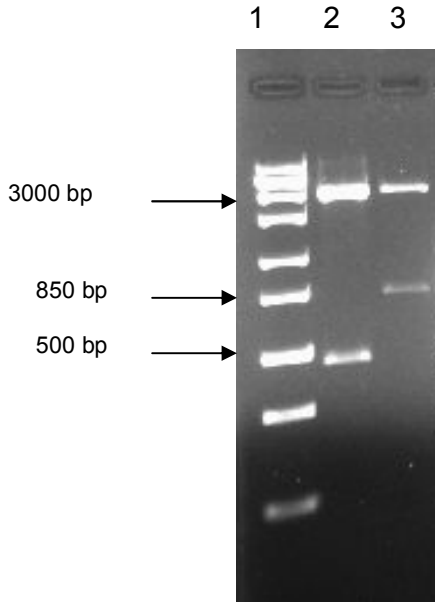


Figure 3.4: Restriction analysis of the PCR product MyrA1.5 cloned into pGEM-T Easy. Purified pGEM-T Easy was digested overnight at 37°C with *EcoR*I. Restriction digests were run on a 1.5% agarose gel. Lane 1: Molecular weight marker, Lane 2: Digested pGEM-T Easy with insert MyrA1.5, Lane 3: Digested pGEM-T Easy with sequence verified 840bp insert

The purified plasmid DNA was sent to Inqaba Biotec for sequencing. Sequencing was done in both directions with the universal SP6 forward and T7 reverse primers. Each sequence was independently aligned to two short pGEM-T Easy sequences either side of the PCR product cloning site using Sequencher®. This was done to identify vector DNA which was trimmed off. The two resulting sequences were aligned. A 448 bp sequence was obtained (data not shown) further indicating that the correct sized PCR fragment had been cloned. Analysis of this sequence however

indicated it was a non-specific product. The sequence was translated in all three frames, with the third frame having the longest ORF. However, there was a single stop codon in this frame. This indicated that the sequence was not correct because both primers were designed from peptide sequences. It would be expected in this case that the fragment would be for an uninterrupted ORF. More important though, neither peptide sequence corresponding to MyrA1 and MyrA5 was present in this sequence. Sequences corresponding to MyrA3 and MyrA7 were also absent.

3.2.7 Amplification of MyrA1.T from cDNA

Amplification with the four forward primers designed for MyrA was attempted with an oligodT primer. Amplification was successful with MyrA1. A fragment of approximately 1.2 kb was obtained (Figure 3.4) hereafter referred to as MyrA1.T. The fragment was amplified in a 50 µl cocktail with 500 ng cDNA, 4 µM of each primer, 2 mM $MgCl_2$, 200 µM of each dNTP and 1.25 U *Taq* polymerase. The cycling conditions were as follows: initial denaturation at 94°C for 2 min, 40 cycles of denaturation at 94°C for 30s, annealing at 45°C for 30s, extension at 72°C for 90s, and final denaturation at 72°C for 5 min..

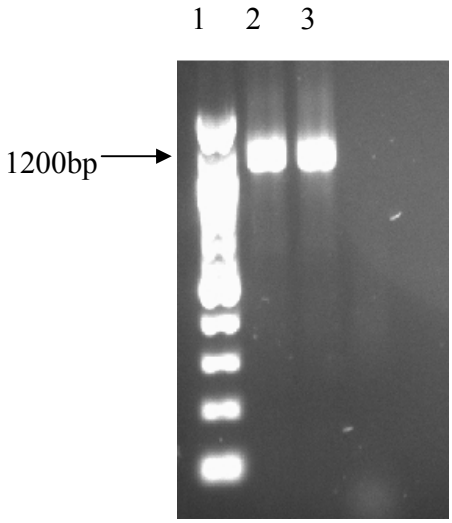


Figure 3.5: Amplification of fragment MyrA1.T. Samples were run on a 1.5% agarose gel. Lane 1: Molecular weight marker; Lanes 2 and 3: MyrA1.T

3.2.8 Cloning and Sequencing of MyrA1.T

Following optimisation of the vector:insert ratio, MyrA1.T was successfully cloned into pGEM-T Easy. While ligation efficiency was low, the presence of an insert was confirmed in a single positive clone designated MyrA1.Tc (Figure 3.6).

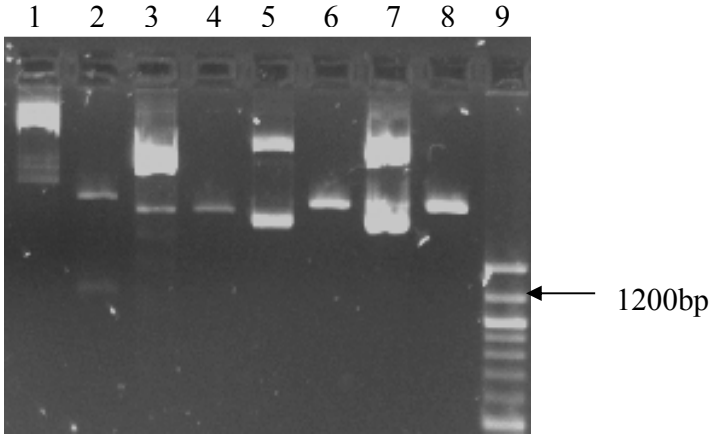


Figure 3.6: Restriction analysis of the PCR product MyrA1.T cloned into pGEM-T Easy. Purified pGEM-T Easy was digested overnight at 37°C with *EcoR*I. Lane 1: MyrA1.Tc undigested, Lane 2: MyrA1.Tc digested, Lane 3: MyrA1.Td undigested, Lane 4: MyrA1.Td digested, Lane 5: MyrA1.Te undigested, Lane 6: MyrA1.Te digested, Lane 7: MyrA1.Tf undigested, Lane 8: MyrA1.Tf digested, Lane 9: Marker

Purified plasmid DNA was sent to Inqaba Biotec for sequencing using the SP6 forward primer. This was because the presence of a polyA tail at one end of the fragment resulted in "dirty" sequence when sequencing was attempted in the reverse direction.

An initial 745 bp sequence was obtained. An internal sequencing primer (5'-CGAAGCATGGATTCAGGTGGTGA-3') starting at position 552 of this sequence was designed to produce the remainder of the sequence. Removal of vector sequence from this sequencing reaction produced an additional 496 bp sequence. Alignment of the two sequences produced a 1047 bp consensus sequence. This sequence was translated in all three frames. The correct ORF was found to begin in the first frame. The peptide fragment corresponding to MyrA1 was identified, beginning at position 1 of the translated sequence indicating that this was the correct clone.

3.3 Sequence analysis of MyrA1.T, a partial sequence for Myr A

The 1047 bp sequence corresponding to MyrA1.T is indicated in Figure 3.7. An 882 bp ORF beginning at position 1of the cDNA sequence was identified. This ORF encodes a 294 amino acid sequence indicated in Figure 3.8. A 147 bp 3' untranslated region follows before the polyA tail.

At least five short peptide sequences (MyrA1, MyrA2, MyrA3, MyrA5 and MyrA6) were positively identified in the predicted amino acid sequence. As mentioned previously, mass spectrometry cannot distinguish between amino acids of similar size. Thus the sequences identified in Figure 3.8 are not identical to those given in Table 3.2. In this instance, isoleucine was sometimes replaced with leucine and vice versa. A sequence matching seven of the nine amino acids in the peptide MyrA7 was identified. MyrA4 was not identified and it is possible that it occurs upstream of MyrA1.

Despite the unsuccessful amplification attempts using primers designed from MyrB fragments, sequences virtually identical to MyrB1, MyrB2 and MyrB3 were identified in the predicted sequence for MyrA (Figure 3.8). These differed by one (MyrB1 and MyrB2) or two (MyrB3) amino acids.

AAACTAGACAACAACCCGAGACATTTCCATCTAGCCGGTAATCCAGA
AGAAGAGTTCCAGCACCAAGGACGGGAAGGGGAACGCAAACCCTCA
ACTCGCGGGGCCTGCAACAATATTTTCTATGGATTGAATGCGAAGAT
CATCGCCGAGGCTCTCAATATTGACGAAAACTTAGCAAGGATACTTA
AAAGTGAGAACGACAACAGAGGCCAGATTGTGAAGGTGAAAGACGG
ACTTCAGGTGGTCGGACCACCCACAAGGCTACAGGAGGAAGAGCGA
GAAGAGGGTGGAAGATCACCACGACGCGGCGAAGGCGACAATGGCG
TTGAGGAGACACTGTGCAACCTCAGGTTAGGAGACAATATCGGTGAT
CCCACACGCGCCGACGTGTTCAACCCAGACGCTGGACGTGTTAGCAT
CCTCAATAGCCAGAACCTACCTGTTCTCCGGGATCTCCGGCTCAGCG
TTGAGGGCGGTGTTCTCTACCACGATGGTATGGTCTTACCCCACTGG
AACAGGAACGCCCACAGCATACTGTACGTGATCAGGGG**CGAAGCATG**
GATTCAGGTGGTGAATGAAGACGGCCAAGCCGTGTTCGACGGCGACT
TGCGCAAGGGGCAGGTGTTGGTCGTGCCACAAAACTTCGCAGTGGTG
AAACGGGCGGAGAAGGAGAGATTCGAATGGGTTGCCTTCAAGACCAA
CGACAAGGCCGCGATCGCACCTCTTGCAGGGCGAACCTCTGCCATTC
GGGCGATGCCTGTGGAAGTAATCGCCAACGCTTACGGGATTTCGTTG
GAGGAAGCAAGGAGGATCAAATTTAACAACAAGCAGACCACTTTGAG
TAGCCCAAGGTCTCAATCCGAAGGATATGCTGCCGATGCT*TGA*AGATTT
TGTGTAAAGATTGCGTAATAATGCAGGCAGACTCTGGTTTTTTGATGTGCAGAA
TAAAAGGAGACTAGTTATTGGTCCTTTTGTTAATGTCTTATGTAATAAGAAGTGT
*ATCACTTAATAATAATAAGCTCTTCTCTCTGC*AAAAAAAAAAAAAAA

Figure 3.7: Sequence of MyrA1.T. Binding position of primers used to amplify the fragment, MyrA1 and the oligodT are highlighted in red. The internal sequencing primer is highlighted in orange. The ORF identified in MyrA1.T is indicated in bold typeface. The stop codon is in italics and underlined. The 3' untranslated region is indicated in italics.

A domain search for MyrA1.T produced a match with the cupin_1 family of proteins characterised by the presence of one or more cupin barrel domains (Dunwell, 1998). The cupin_1 family is a member of the cupin superfamily. Members of the cupin_1 family include the ubiquitous plant seed storage proteins. A pairwise sequence alignment of the Myr A amino acid sequence and the consensus sequence for the cupin 1 family indicating the conserved cupin barrel domain is shown in Figure 3.9. The multiple sequence alignment was generated in NCBI CDD. The location of the two motifs is indicated in Figure 3.8.

KLDNNPRHFHLAGNPEEEFQHQGREGERKPSTRGACNNIFYGLNAKIIAEALN

MyrA1 MyrA3

IDENLARILKSENDNRGQIVKVKDGLQVVGPPTRLQEEEREEGGRSPRRGEGD

MyrA5 MyrA2

NGVEETLCNLRLGDNIGDPTRADVFNPDAGRVSILNSQNLPVLRDLRLSVEG

MyrB2? MyrB3? MyrA6

GVLYHDGMVLPHWNRNAHSILYVIRGEAWIQVVNEDGQAVFDGDLRKGQV

Motif 1

LVVPQNFAVVKRAEKERFEWVAFKTNDKAAIAPLAGRTSAIRAMPVEVIAN

Motif 2 MyrB1? MyrA7?

AYGISLEEARRIKFNNKQTTLSSPRSQSEGYAADA

Figure 3.8: A 294 **amino acid sequence identified in the MyrA1.T cDNA fragments.** The position of the peptide sequence MyrA1 is highlighted in red. Other peptide fragments are highlighted in orange, with MyrA3 also underlined.

The predicted molecular weight of the protein was computed to be approximately 32.6 kDa using the Expasy tool ProtParam. This compares with the estimation of Myr A using SDS-PAGE of approximately 30 kDa.

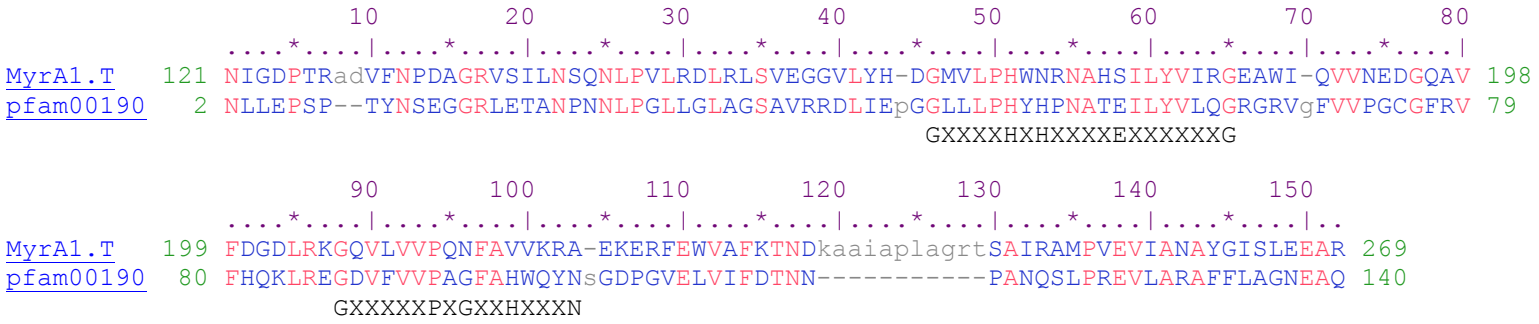


Figure 3.9: Pairwise sequence alignment of the consensus sequence for the conserved cupin barrel domain of cupin 1 family and MyrA1.T. The pairwise sequence alignment was generated by querying the NCBI CDD database for conserved domains in the Myr A sequence. The match was significant with an E-value of 2e-25 In general, residues highlighted in red are highly conserved and those in blue less conserved. Pfam00190 = cupin 1 family.

3.3 Identification of sequences similar to MyrA1.T

3.3.1 Identification of nucleic acids sequences similar to MyrA1.T

BLAST searches were done with the amino acid and cDNA sequences to find known sequences similar to MyrA1.T. Blastn was used to probe the GenBank database for similar nucleic acid sequences. The result of the BLASTn search is indicated in table 3.6, with the top 10 matches indicated.

MyrA1.T showed the most similarity to a predicted protein *Populus trichocarpa* mRNA with an E-value of $9e^{-110}$. Myr1A1.T was aligned to 71% of this sequence with an identity of 72%. Other sequences identified corresponded to mRNA coding for seed storage proteins in other species including *Ricinus communis* alegumin B precursor mRNA (E-value: $3e^{-78}$) and *Ficus pumila* var. awkeotsang 11S globulin precursor isoform 3A mRNA (E-value: $3e^{-53}$)

Table 3.6: Nucleic acid sequences matching MyrA1.T. Sequences were identified using the blastn algorithm and are filtered according to their E value ranked from lowest to highest

Accession	Description	Query coverage	% Identity	E value
XM_002300739.1	*Populus trichocarpa* predicted protein, mRNA	71%	72%	$9e^{-110}$
XM_002307609.1	*Populus trichocarpa* predicted protein, mRNA	71%	69%	$4e^{-82}$
XM_002524152.1	*Ricinus communis* legumin B precursor, putative, mRNA	71%	69%	$3e^{-78}$
AF262998.1	*Ricinus communis* legumin-like protein mRNA, complete cds	71%	69%	$3e^{-78}$
XM_002524560.1	*Ricinus communis* legumin A precursor, putative, mRNA	71%	68%	$2e^{-68}$
XM_002524149.1	*Ricinus communis* legumin B precursor, putative, mRNA	71%	68%	$2e^{-68}$
XM_002531982.1	*Ricinus communis* legumin A precursor, putative, mRNA	67%	68%	$2e^{-60}$
XM_002524558.1	*Ricinus communis* legumin A precursor, putative, mRNA	71%	67%	$3e^{-59}$
XM_002524559.1	*Ricinus communis* legumin A precursor, putative, mRNA	71%	67%	$3e^{-58}$
EF091697.1	*Ficus pumila* var. awkeotsang 11S globulin precursor isoform 3A mRNA, complete cds	49%	71%	$3e-^{53}$

3.3.2 Identification of amino acid sequences similar to MyrA1.T

Blastp was used to probe the GenBank database for sequences similar to the predicted MyrA1.T amino acid sequence. The most similar sequence identified was a predicted protein from *P. trichocarpa*, with an E-value of $1e^{-81}$ and query coverage of 99%. All other significant matches were seed storage proteins and related proteins. The top 10 matches are indicated in table 3.6.

Table 3.7: Amino acid sequences matching MyrA1.T. Sequences were identified using the blastp algorithm and are filtered according to their E value ranked from lowest to highest.

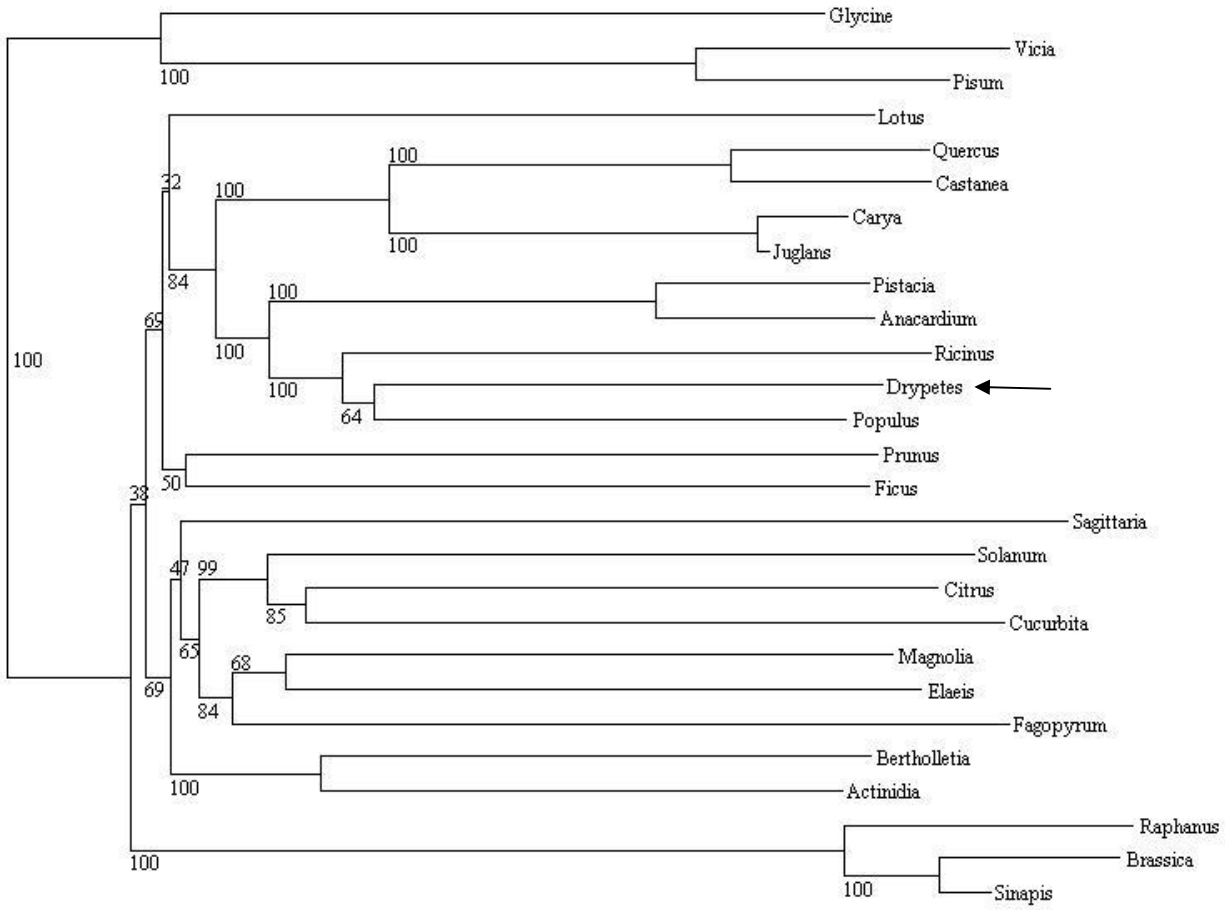
Accession	Description	E value
XP_002307645.1	predicted protein (*Populus trichocarpa)*	$1e^{-97}$
XP_002300775.1	predicted protein (*Populus trichocarpa)*	$1e^{-93}$
XP_002524606.1	legumin A precursor, putative (*Ricinus communis*)	$4e^{-86}$
XP_002524605.1	legumin A precursor, putative (*Ricinus communis*)	$1e^{-85}$
XP_002524195.1	legumin B precursor, putative (*Ricinus communis*)	$1e^{-82}$
XP_002524604.1	legumin A precursor, putative (*Ricinus communis*)	$4e^{-82}$
XP_002524198.1	legumin B precursor, putative (*Ricinus communis*)	$3e^{-81}$
ACB55490.1	Pis v 5.0101 allergen 11S globulin precusor (*Pistacia vera*)	$5e^{-80}$
ABI94736.1	11S seed storage globulin B (*Chenopodium quinoa*)	$6e^{-80}$
AAS67036.1	11S seed storage globulin (*Chenopodium quinoa*)	$1e^{-79}$

3.4 Phylogenetic analysis of MyrA1.T

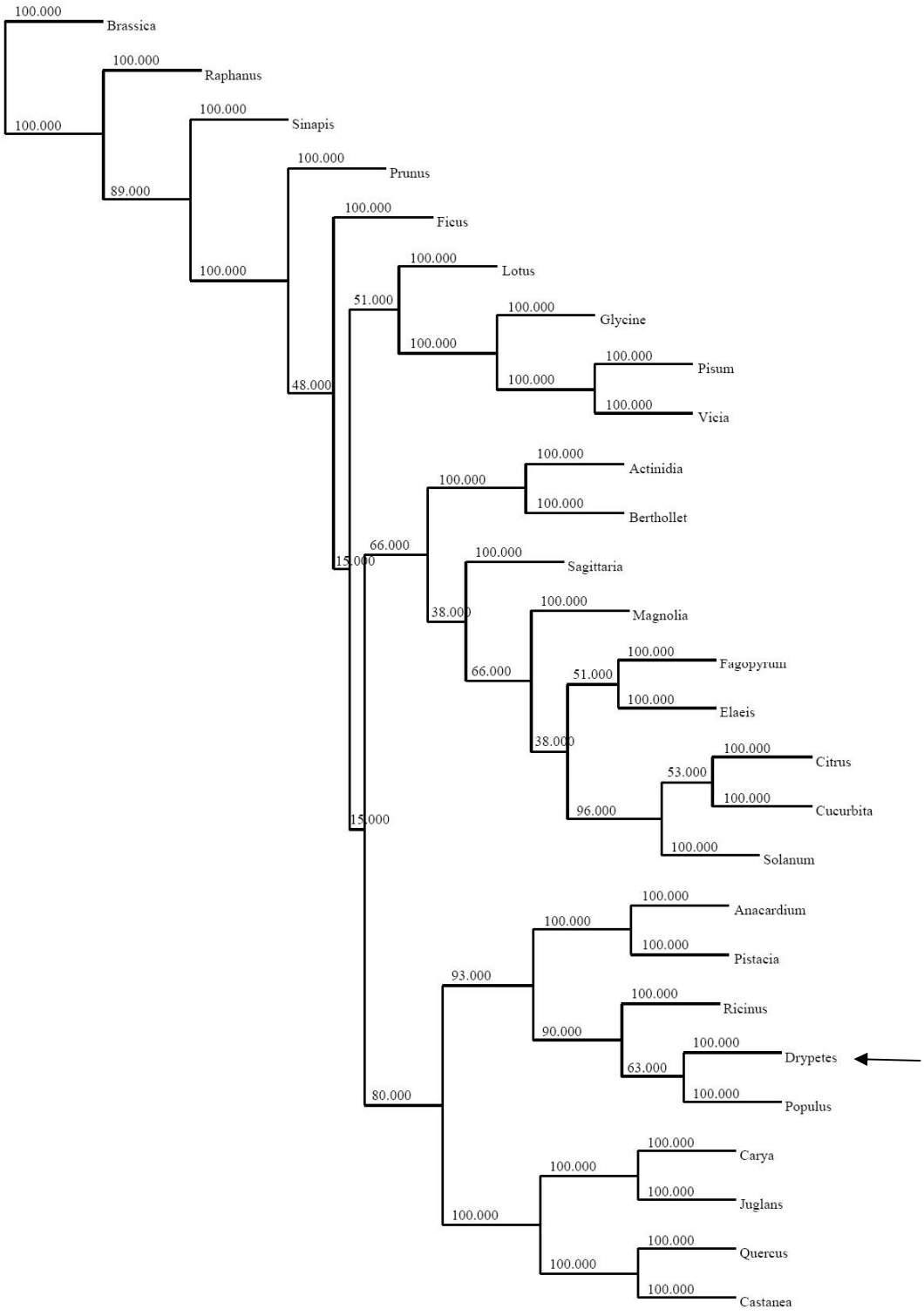
MyrA1.T was aligned to 26 mRNA sequences that showed significant similarity using ClustalX. The phylogenetic tree constructed using the neighbour-joining method was computed using ClustalX with bootsrapping with 100 replicates. Maximum-likelihood and maximum-parsimony methods were computed using the neighbour, Dnaml and Dnapars programs respectively from the PHYLIP package. For each method, bootstrap analysis was done with 100 replicates using the program Seqboot, and a consensus tree was computed using the program Consense.

The three methods produced trees with similar but slightly differing topologies (Figure 3.10). In all three trees, MyrA1.T formed a cluster with sequences from *R. communis*, *P. trichocarpa*, *P. vera* and *Anacardium occidentale* with bootstrap support of 100% for neighbour-joining, 93% for maximum-likelihood analysis and 92% maximum-parsimony analysis. In neighbour-joining analysis (Figure 3.10 (c)) and maximum-likelihood analysis (Figure 3.10 (b)), this formed a larger cluster with sequences from *Carya illinoinensis*, *Juglans regia*, *Quercus robur* and *Castanea crenata* with bootstrap support of 84% and 80% respectively. In maximum parsimony analysis (Figure 3.10 (b)) this differed, with the MyrA1.T cluster more related to sequences from *Ficus pumila* and *Prunus amygdalus* albeit with low bootstrap support (26%). The cluster with sequences from *Carya illinoinensis*, *Juglans regia*, *Quercus robur* and *Castanea crenata* was grouped with sequences from *S. alba*, *B. napus*, *R.sativus, Lotus japonica, Pisum sativum*, *Glycine max* and *Vicia faba* with bootstrap support of 31.6%.

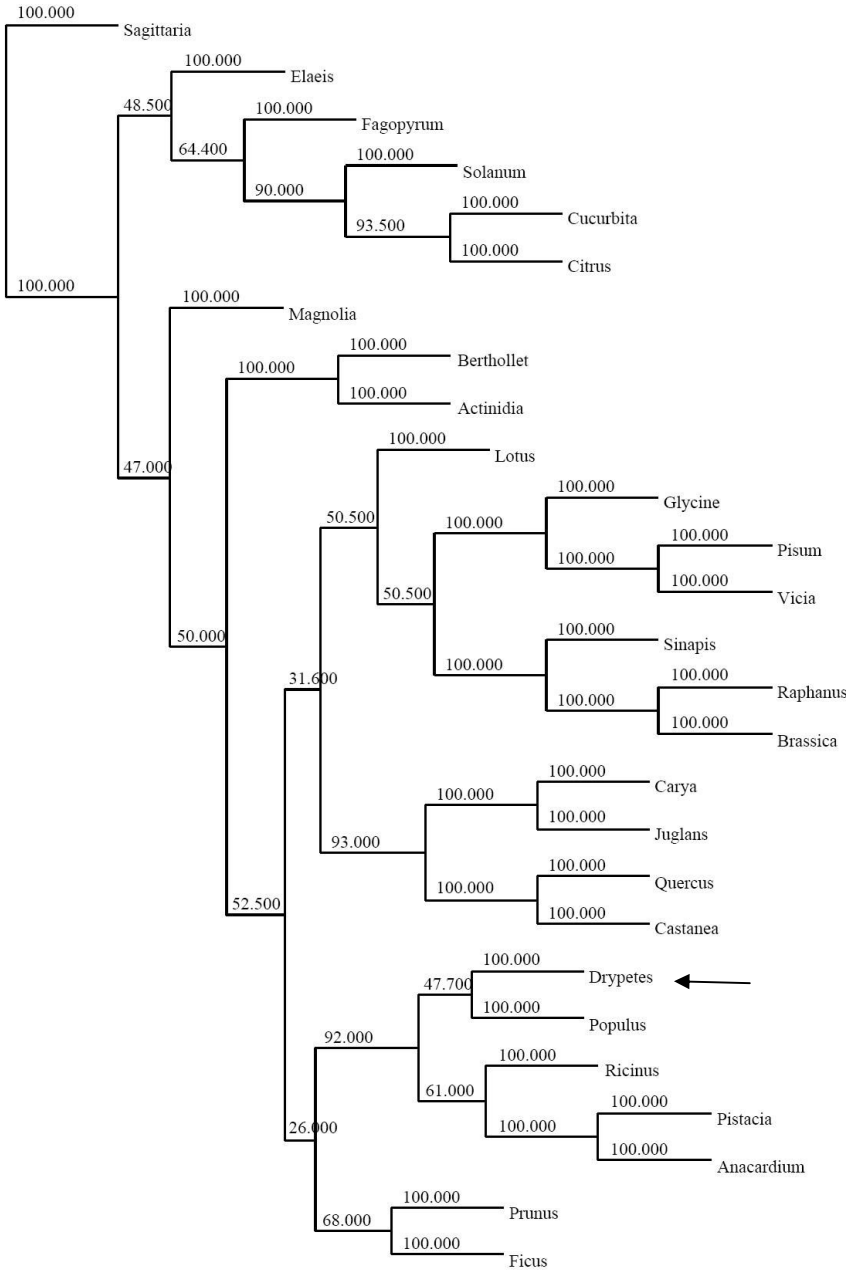
.


Glycine
Vicia
Pisum
Lotus
Quercus
Castanea
Carya
Juglans
Pistacia
Anacardium
Ricinus
Drypetes
Populus
Prunus
Ficus
Sagittaria
Solanum
Citrus
Cucurbita
Magnolia
Elaeis
Fagopyrum
Bertholletia
Actinidia
Raphanus
Brassica
Sinapis
100
32
84
69
38
47
99
85
65
68
64
50

(a)


100.000
Brassica
100.000
Raphanus
100.000
Sinapis
100.000
Prunus
100.000
Ficus
100.000
Lotus
100.000
Glycine
100.000
Pisum
100.000
Vicia
100.000
Actinidia
100.000
Berthollet
100.000
Sagittaria
100.000
Magnolia
100.000
Fagopyrum
100.000
Elaeis
100.000
Citrus
100.000
Cucurbita
100.000
Solanum
100.000
Anacardium
100.000
Pistacia
100.000
Ricinus
100.000
Drypetes
100.000
Populus
100.000
Carya
100.000
Juglans
100.000
Quercus
100.000
Castanea
89.000
100.000
51.000
48.000
15.000
66.000
38.000
66.000
51.000
38.000
53.000
96.000
15.000
93.000
90.000
63.000
80.000
100.000
100.000
100.000

(b)

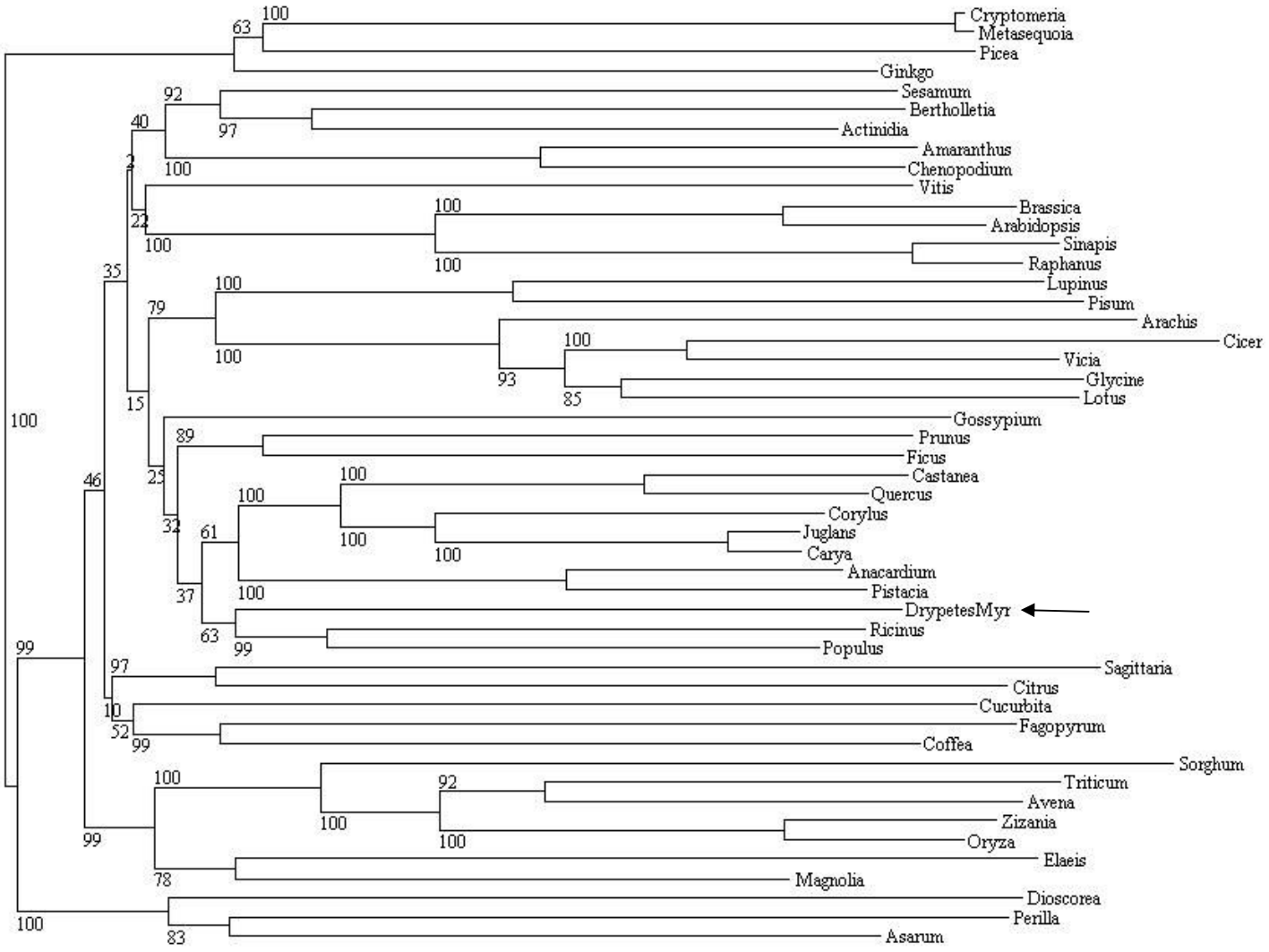


(c)

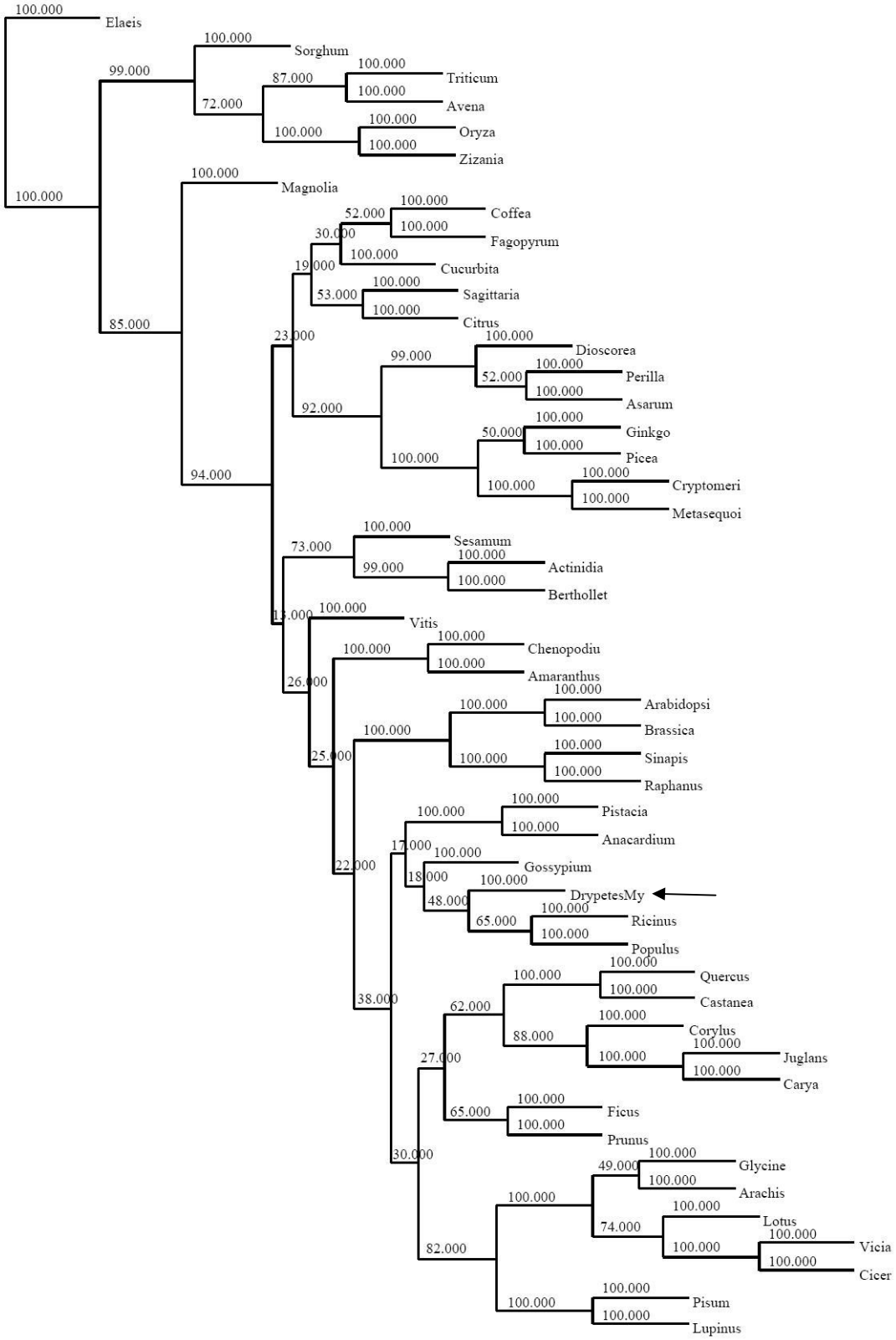
Figure 3.10: **Phylogenetic trees indicating the relationship between MyrA1.T cDNA and related sequences.** Bootstrap support is indicated above the branches. (a) Neighbour-joining (b) Maximum-likelihood (c) Maximum-Parsimony. Arrows indicate the position of *Drypetes*

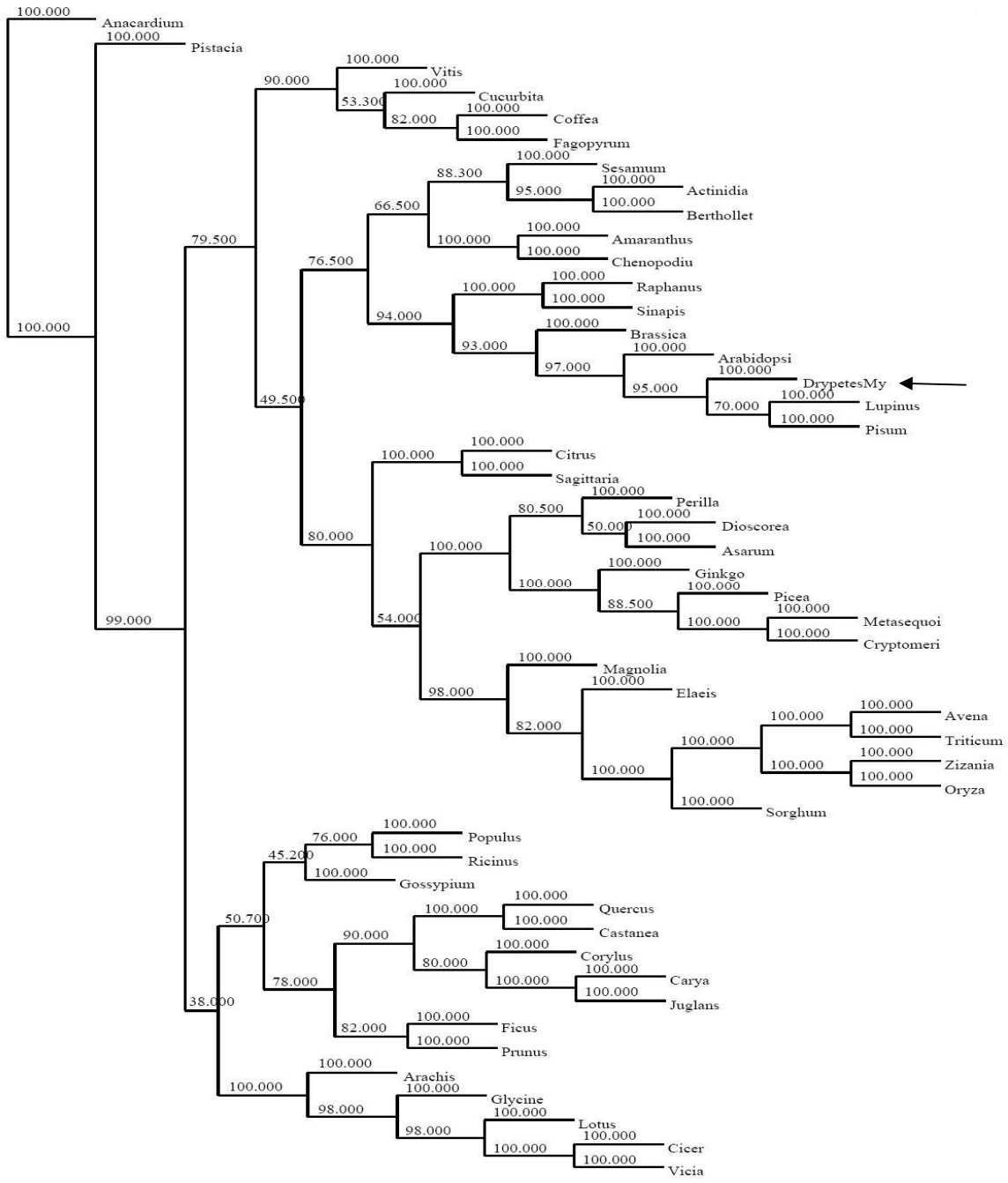
Forty eight similar sequences were aligned to the predicted amino acid sequence of MyrA1.T (Figure 3.11). Phylogenetic trees constructed using the neighbour-joining method were computed with ClustalX with bootstrapping for 100 replicates. Maximum-likelihood and maximum-parsimony were computed using the protml and protpars programs respectively from the PHYLIP package. Similar to nucleotide analysis, for each method, bootstrap analysis was done with 100 replicates using the program seqboot, and a consensus tree was computed using the program consense.

Using neighbour-joining analysis *Drypetes* MyrA1.T was found in a small cluster with *R. communis* and *P. trichocarpa* with 63% bootstrap support. This was found in a larger cluster with sequences from *C. illinoinensis*, *J. regia*, *Q. robur* and *C. crenata*, *Corylus avellana P. vera* and *Anacardium occidentale*. This cluster however had a low bootstrap support of 37%. Like with results obtained using maximum-likelihood analysis with nucleotides, *Drypetes* MyrA1.T clustered with *R. communis*, *P. trichocarpa*, *P. vera* and *Anacardium occidentale*, and additionally a sequence from *Gossypium hirsutum* for which this was no corresponding nucleotide sequence. This cluster however had a low bootstrap support of 17%. Maximum-parsimony produced a topology distinct from neighbour-joining and maximum-likelihood. MyrA1.T formed a cluster with sequences from *Pisum sativum* and *Lupinus angustifolius* with bootstrap support of 95%. This was further found in larger cluster with sequences from *Arabidopsis, B. napus, S. alba* and *R. sativus* which belong to the order Capparales, with bootstrap support of 94%


Cryptomeria
Metasequoia
Picea
Ginkgo
Sesamum
Bertholletia
Actinidia
Amaranthus
Chenopodium
Vitis
Brassica
Arabidopsis
Sinapis
Raphanus
Lupinus
Pisum
Arachis
Cicer
Vicia
Glycine
Lotus
Gossypium
Prunus
Ficus
Castanea
Quercus
Corylus
Juglans
Carya
Anacardium
Pistacia
DrypetesMyr
Ricinus
Populus
Sagittaria
Citrus
Cucurbita
Fagopyrum
Coffea
Sorghum
Triticum
Avena
Zizania
Oryza
Elaeis
Magnolia
Dioscorea
Perilla
Asarum

(a)


Elaeis
Sorghum
Triticum
Avena
Oryza
Zizania
Magnolia
Coffea
Fagopyrum
Cucurbita
Sagittaria
Citrus
Dioscorea
Perilla
Asarum
Ginkgo
Picea
Cryptomeri
Metasequoi
Sesamum
Actinidia
Berthollet
Vitis
Chenopodiu
Amaranthus
Arabidopsi
Brassica
Sinapis
Raphanus
Pistacia
Anacardium
Gossypium
DrypetesMy
Ricinus
Populus
Quercus
Castanea
Corylus
Juglans
Carya
Ficus
Prunus
Glycine
Arachis
Lotus
Vicia
Cicer
Pisum
Lupinus



(c)

Figure 3.11: Phylogenetic trees indicating the relationship between the MyrA1.T amino acid sequence and related sequences. Bootstrap support for 100 replicates is indicated. (a) Neighbour-joining (b) Maximum-likelihood (c) Maximum-parsimony. Arrows indicate the position of *Drypetes.*

4. DISCUSSION

Myrosinase activity was detected in one species, *Drypetes natalensis.* The candidate protein with myrosinase activity was found to be a 50 kDA heterodimer with subunits of approximately 30 kDA (Myr A) and 20 kDa (Myr B). This contrasts with other myrosinases, which were found to occur as homodimers with subunits of between 65 and 75 kDA (Rask et al, 2000). A non-plant myrosinase from the cabbage aphid *B. brassicae* was detected as a homodimer with subunits of between 51 kDa and 57 kDa according to different sources **(**Husebye et al, 2005, Jones et al, 2001, Pontoppidan et al, 2001).

Sequencing of the candidate myrosinase in *Drypetes* via tandem mass spectrometry produced nine short peptide sequences and three short peptide sequences for Myr A and Myr B respectively. As expected, neither subunit sequence matched the sequences of any known myrosinase genes. Sequences for Myr A produced non-significant matches to a legumin-like protein from *Ricinus communis* while sequences for Myr B matched a putative 11S seed storage protein. Coupled with the results obtained from the characterisation of myrosinase, this is suggestive of an independent evolution of myrosinase activity in *D. natalensis* from a novel ancestral source.

A partial coding region for the larger subunit of the candidate myrosinase was successfully isolated using a low stringency degenerate primer PCR strategy. The results obtained were in accordance with the identification from the mass spectrometry data. Both the MyrA1.T cDNA and the predicted amino acid sequence showed significant similarity to the ubiquitous seed storage proteins from a number of species and not to any known myrosinases.

Five of the short peptide sequences derived from MyrA (MyrA1, MyrA2, MyrA3, MyrA5 and MyrA6) were identified in this coding sequence. A sequence

corresponding to a portion of MyrA7 was also identified, differing from Maldi-TOF with the first two amino acids. MyrA4 was not identified and this sequence probably occurs upstream of MyrA1. Furthermore, three sequences differing by one (MyrB1 and B2) or two (MyrB3) amino acids, from the short peptide sequences obtained for MyrB were identified in the MyrA sequence. This suggests that Myr is formed from the products of two related genes that associate to form an active heterodimer. This is based primarily on the observed differences, though small, in sequences obtained from short peptide sequences compared to those identified on the predicted sequence. While these differences could be due to errors in peptide sequencing, of note is that due to the degeneracy of the genetic code a small difference in amino acid sequence is far greater at the nucleotide level. This could suggest that MyrB is encoded by a gene related to *MyrA*. This could have arisen through a duplication event followed by sequence degradation to yield a truncated form of MyrA.

The phylogeny of Myr A and related sequences was assessed at the amino acid and nucleic acid sequence level using three methods, neighbour-joining, maximum parsimony and maximum likelihood. There was overall a broad similarity between the results obtained using the different methods, with *Drypetes* Myr A consistently branching with sequences from *R. communis* and *P. trichocarpa*. Figure 4.1 (a) and (b) are phylogenetic trees depicting the phylogeny of angiosperms and the order Malpighiales respectively (Stevens, 2001). *R.* communis and *P. trichocarpa* like *Drypetes* are of the order Malpighiales, in the families Euphorbiaceae and Salicacae respectively. *Drypetes* belongs to the family Puntranjivaceae. This explains why *Drypetes* MyrA consistently clustered with these sequences irrespective of the type of analyses being performed. Most sequences that were in close proximity to MyrA were rosids and asterids. Sequences that were more distantly related to MyrA were monocots, belonging to the group's commelinids or magnolids, with the exception of the sequence for *Magnolia.* Again this can be explained, as MyrA has clustered with sequences belonging to related eudicots as opposed to more distantly related monocots.

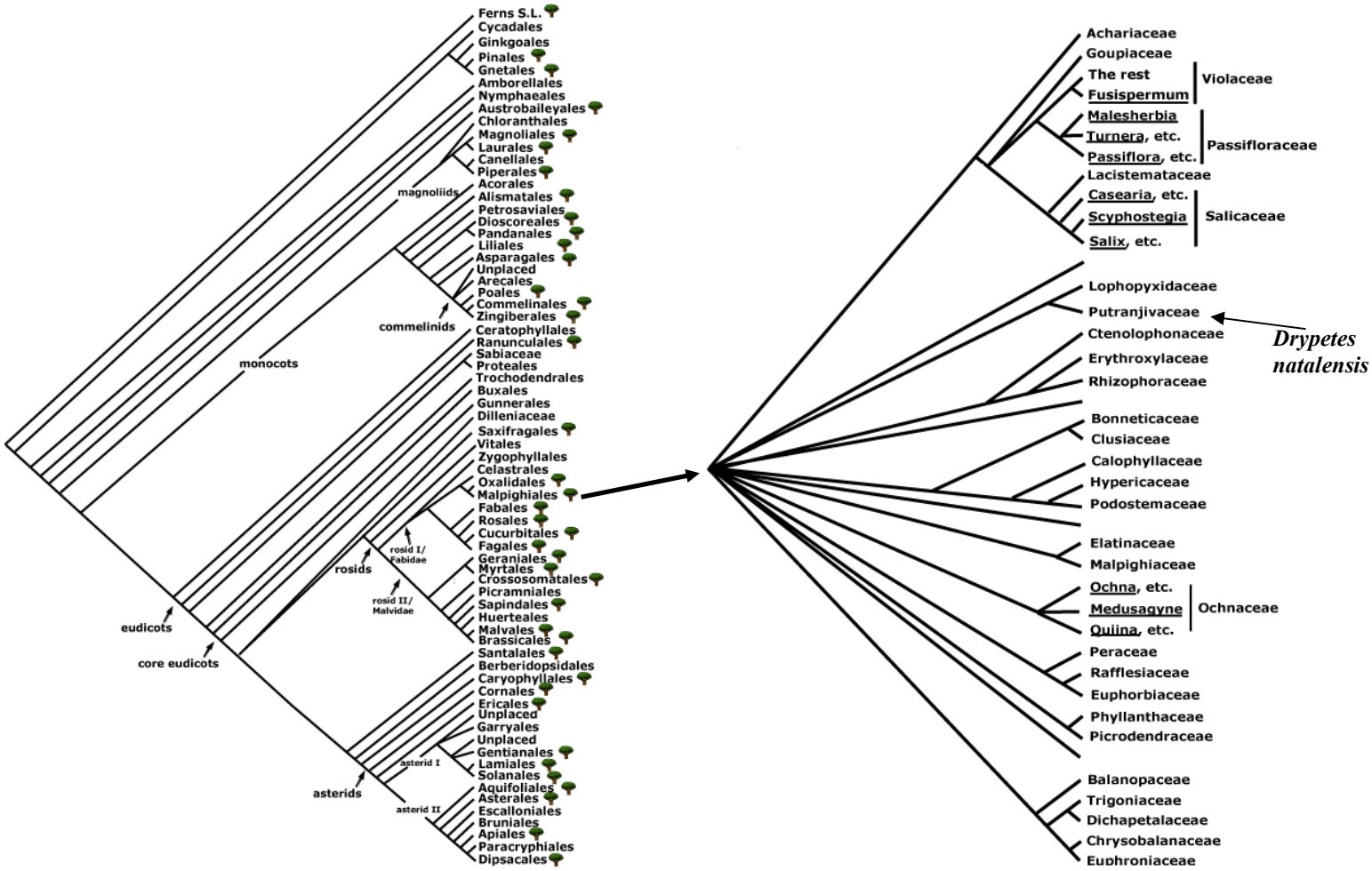


Figure 4.1: Trees indicating the phylogeny of (a) the angiosperms and (b) the order Malpighiales
(adapted from Stevens (2001))

Despite the successful detection of myrosinase activity, and the isolation of a candidate molecule, some questions remain. Time and resource constraints unfortunately meant that purification of the candidate enzyme was not possible. This would have allowed for further characterisation. Several parameters already determined for myrosinases, like optimum temperature and pH values, could not be directly compared to myrosinase activity in *Drypetes.* Furthermore, it is well established, for example, that myrosinases require ascorbic acid as a co-factor for optimal activity (Rask et al, 2000). Even in non-plant systems, contradictory results have been observed with one study showing an inhibitory effect (Pontoppidan et al, 2001) and another showing no effect on enzyme activity (Husebye, 2005). As the enzyme responsible for myrosinase activity in *Drypetes* was not purified to homogeneity it was not possible to draw any conclusions in this regard. As it is impossible to speculate on the effect of ascorbic acid on *Drypetes* myrosinase, further work is needed to clarify if ascorbic acid has any effect, as this seems to be important in the functioning of plant myrosinases.

Also important to note is that non-denaturing gel assays (Figure 3.1) showed that myrosinase activity migrated with a highly abundant protein. That sequencing produced a coding sequence with significant similarity to seed storage proteins does raise some doubts about the validity of the isolated sequence being the most likely candidate. Seed storage proteins, mostly seen as nutrient reservoirs and possessing no enzymatic activity, are the highly abundant proteins in growing seedlings (Shewry et al, 1995). In maize, a group of storage proteins called zeins make up 50% of the protein content of the seed (Chui and Falco, 1995). In soybean, the globulins account for 70-80% of the total seed protein content (Meinke et al, 1981; Krishnan et al, 2000) and 30-40% of the dry weight of the seed (Meinke et al, 1981). Enzymes are conversely found in very low quantities, with very little of the catalyst required for catalytic activity. Thus it cannot be discounted that the factor responsible for myrosinase activity co-migrated with the isolated seed storage protein and was simply found in quantities below the level of detection of PAGE and specifically the

use of the Coomassie brilliant blue stain in SDS-PAGE analysis. Again, purification and concentration of the enzyme to homogeneity would have been preferred to determine if the candidate protein was responsible for the observed activity.

The protein isolated here like the seed storage proteins belongs to the cupin superfamily as indicated by the significant match to the cupin domain when probing the NCBI CDD. The cupin superfamily is considered to be one of the most diverse superfamilies recorded (Dunwell et al, 2004) with at least 18 defined subclasses. It has been shown to include both enzymatic and non-enzymatic members (Dunwell et al, 2000). Seed storage proteins are members of the cupin 1 family of proteins within the cupin superfamily (Dunwell, 1998). This family is characterised by the presence of the cupin domain first identified in the wheat protein germin. The cupin domain is comprised of two conserved motifs separated by a divergent region varying in length from as little as 10 amino acids to more than 100 amino acids (Dunwell et al, 2004). Each motif forms a β-sheet, folding into a barrel shape hence their designation as cupins. The cupin domain is ubiquitous and has been detected in animals, plants, fungi and prokaryotes (Dunwell et al, 2000).

Cupins can be found as either monocupins or in a multidomain configuration (Dunwell et al, 2004). Most enzymatic members of the cupin superfamily are monocupins with the domain found at the centre of the protein however several enzymes also possess the bicupin configuration. (Dunwell et al, 2000). The first identified cupin, the wheat protein germin, is an oxalate oxidase involved in the formation of CO_2 and hydrogen peroxide from oxalate (Dunwell et al, 2004). Other enzymatic monocupins are phosphomannose isomerase and dioxygenases (Dunwell et al, 2000). Polyketide synthases, enzymes involved in the biosynthesis of antibiotics, are also monocupins. Auxin-binding proteins have also been shown to possess single cupin domains. Seed storage proteins fall into the class of bicupins, which also includes a number of prokaryotic proteins (Dunwell et al, 2000). The most closely related bicupins to the seed storage proteins are the sucrose-binding proteins.

The two conserved motifs found in the cupin domain have a consensus sequence of $G(X)_5\,HXH(X)_{3,4}E(X)_6G$ for motif 1 and $G(X)_5PXG(X)_2H(X)_3N$ for motif 2. These sequences were found in Myr A though some residues were not conserved. It has been shown by analysing several cupin sequences that this is often the case (Dunwell et al, 2004). MyrA lacks the first G (replaced with D), the second H (N) and the E (S) in motif 1. In motif 2 the second G has been replaced with N, the H with V and N replaced with A (Figure 3.9). These motifs were identified in the wheat germin protein with some of the residues, most notably the G and the H in motif 1 and H in motif 2 critical for the germin enzymatic activity (Dunwell et al, 2004). These residues are involved in the binding of manganese which is required for the activity of germin (Woo et al, 2000). Comparison of its structure to two non-catalytic relatives showed that they lacked the G and at least one of the H residues, suggesting that seed storage proteins are deactivated enzymes (Dunwell et al, 2004).

The isolated Myr differs from seed storage proteins in that it appears to be a monocupin while still lacking some of the critical amino acids thought be required for enzymatic activity in this superfamily. This points towards a non enzymatic function but from results obtained here Myr could possess myrosinase-like activity and be the factor that is able to hydrolyse the glucosinolate sinigrin. Is it possible that Myr is a modified seed storage protein that under selective pressure has acquired glucosidase activity similar to that in the Capparales, much like in aphids and other non plant systems?

Aphid myrosinase, though related to β-glucosidases was shown to be more related to animal glucosidases than to plant myrosinases or plant glucosidases both at a sequence level and at a structural level (Jones et al, 2002; Husebye et al, 2005) This is reflected in the architecture of its active site. While aphid myrosinase has a similar fold in the active site, key residues critical for substrate recognition and hydrolysis, differ from plant myrosinase (Husebye et al, 2005). What is interesting about aphid

myrosinase is that even though the active site of plant myrosinases diverged from that of other β-glucosidases to accommodate and hydrolyse glucosinolates, it has remained largely unchanged in aphid myrosinase. Thus aphid myrosinase evolved in such a manner that it is able to hydrolyse glucosinolates without adopting similar active site changes to accommodate a different substrate.

A recent study by Gherardini et al (2007) showed evidence for the common occurrence for convergent evolution in the active sites of enzymes. The authors were looking to establish how different enzymes can adapt to solve a common catalytic conundrum. They grouped convergently evolved enzymes into mechanistic and transformational analogues. At a simplistic level, transformational analogues are non-homologous enzymes which effect the same change, as defined by enzyme commission (EC) numbers, usually by different means, whereas mechanistic analogous share common catalytic residues, but can also have different substrate specificities . There can be overlaps, where an enzyme is both a transformational and mechanistic analogue. Mechanistic analogues were common, more so than transformational analogues, occurring at rates of ~15% and ~6% respectively. Of interest to this research is the phenomenon of transformational analogues as this is the most likely manner of adaptation that has occurred.

It is also generally accepted that tertiary structure is more conserved than amino acid sequences (Grishin, 2001) and in some respects is more informative. This is because there are very few folds that proteins can conform to; hence unrelated sequences can often have similar structural folds which can be detected even where there is little or no similarity between sequences (Todd et al, 1999, Grishin, 2001). This is in turn leads to the possibility of diverse proteins performing similar, analogous functions as was shown by Gherardini et al (2007). Another relevant feature of protein folds is that while there is a high level of conservation, it is also possible within a family, for small sequence changes to effect substantial structural variation within the family which leads to functional diversification (Kinch and Grishin, 2002).

As mentioned above, it is apparent that the cupin domain is important in a number of different biological functions whether enzymatic or non-enzymatic. Of interest though is the functional diversity of the cupins at the enzymatic level. In terms of enzyme function, the cupin superfamily has a high level of plasticity. Some of the work presented here reveals a possible mechanism for the evolution of myrosinase activity in *Drypetes*.

If the isolated candidate protein is indeed a myrosinase-like enzyme, then it is proposed that myrosinase activity in *Drypetes* arose following this likely sequence of events. A series of duplication events first resulted in two or possibly more copies of a gene encoding a seed storage protein. Increased sequence diversion resulted in related sequences of variable length. In some cases this would have resulted in the loss of the bicupin configuration explaining the presence of single cupin domain in the MyrA sequence. With increased selective pressure, most likely increased herbivory, mutations that resulted in gene products able to hydrolyse glucosinolates, where hydrolysis product would act as a deterrent, were fixed. It is unclear if pathways involved in the biogenesis of the required glucosinolates would have been present in *Drypetes* at this stage or arose following selective pressure. Also not clear is why the active enzyme was isolated as a heterodimer. It is unclear which subunit is responsible for the observed activity, or whether one or both are required for catalytic activity, with the active site present at the dimer interface. Another possibility is that one subunit is required for the stability of the catalytically active subunit. It is also possible that one subunit could be required for substrate binding and positioning, with catalytic activity localised to the second subunit. To answer these questions would require more experimental work which was beyond the aims and scope of this research project.

5. CONCLUSION AND FUTURE WORK

Evidence presented here confirms the presence of myrosinase activity in *Drypetes* and suggest that it was a result of convergent evolution. A 50 kDa heterodimeric candidate protein was isolated using a combination of a native PAGE assay and SDS-PAGE. Sequencing and phylogenetic analysis revealed that it was related to the ubiquitous plant seed storage proteins. The cupin domain, a feature of seed storage proteins but also present in a number of enzymes, was identified in the predicted amino acid sequence. While these proteins are usually non-enzymatic a possible pathway for the evolution of myrosinase activity was postulated.

However, despite some evidence for the evolution of myrosinase activity here, there are some unanswered questions. It is noted that it is possible the enzyme responsible for myrosinase was not detected, and was merely co-migrating with an abundant seed storage protein. To resolve this, purification of the active myrosinase, not possible in this study due to both time and resource constraints, would be the next step. Of great importance is to obtain the full mRNA and genomic sequences. This will allow for further characterisation of myrosinase enabling a thorough comparison with plant myrosinases and aphid myrosinase.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. And Lipma, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410

Andreasson, E., Taipalensuu, J., Rask, L. and Meijer, J. (1999) Age-dependent wound induction of a myrosinase associated protein from oilseed rape (*Brassica napus*). *Plant Molecular Biology* **41**, 171-180

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marchall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., Eddy, S. R. (2004) The Pfam protein families database. *Nucleic Acids Research* **32**, 138-141

Beekwilder, J., van Leeuwen, W., van Dam, N. M., Bertossi, M., Grandi, V., Mizzi, L., Soloviev, M., Szabados, L., Molthoff, J. W., Schipper, B., Verbocht, H., de Vos, R. C. H., Morandini, P., Aarts, M. G. M. And Bovy, A. (2008) The Impact of the Absence of Aliphatic Glucosinolates on Insect Herbivory in Arabidopsis. *PLoS ONE* **3**, e2068

Bernadi, R., Negri, A., Severino, R. and Palmieri, S. (2000) Isolation of the epithiospecifier protein from oil-rape (*Brassica napus* ssp. *oleifera*) seed and its characterization. *FEBS Letters* **467**, 296-298.

Birnboim, H. C. and Doly. J. (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Research* **7,** 1513-1523

Bones A. M. and Rossiter, J. T. (1996) The myrosinase-glucosinolate system, its organization and biochemistry. *Physiologia Plantarum* **97,** 194-208.

Bones, A. M. and Slupphaug, G. (1989) Purification, characterization and partial amino acid sequencing of *β*-thioglucosidase from *Brassica napus* L. *Journal of Plant Physiology* **134**, 722-729.

Bonnesen, C., Eggleston, I. M. and Hayes, J. D. (2001) Dietary indoles and isothiocyanates that are generated from cruciferous vegetables can both stimulate apoptosis and confer protection against DNA damage in human colon cell lines. *Cancer Research* **61,** 6120-6130.

Bourderioux, A., Lefoix, M., Gueyrard, D., Tatibouët, A., Cottaz, S., Arzt, S., Burmeister, W. P. and Rollin, P. (2005) The glucosinolate-myrosinase system. New insights into enzyme-substrate interactions by use of simplified inhibitors. *Organic and Biomolecular Chemistry* **3**, 1872-1879.

Bridges, M., Jones, A. M. E., Bones, A. M., Hodgson, C., Cole, R., Bartlett, E., Wallsgrove, R., Karapapa, V. K., Watts, N. and Rossiter, J. T. (2002) Spatial organization of the glucosinolate-myrosinase system in brassica specialist aphids is similar to that of the host plant. *Proceedings of the Royal Society of London B* **269**, 187-191

Burow, M., Bergner, A., Gershenzon, J. and Wittstock, U. (2007) Glucosinolate hydrolysis in *Lepidium sativum*—identification of the thiocyanate-forming protein. *Plant Molecular Biology* **63,** 49-61

Burmeister, W. P., Cottaz, S., Driguez, H., Iori, R., Palmieri, S. and Henrissat, B. (1997) The crystal structures of *Sinapis alba* myrosinase and a covalent glcosyl-enzyme intermediate provide insights into substrate recognition and active-site machinery of an *S*-glycosidase. *Current Biology* **5**, 663-675.

Burmeister, W. P., Cottaz, S., Rollin, P., Vasella, A. and Henrissat, B. (2000) High-resolution X-ray crystallography shows that ascorbate is a cofactor for myrosinase and substitutes for the function of the catalytic base. *Journal of Biological Chemistry* **275**, 39385-39383.

Chen, S. and Andreasson, E. (2001) Update on glucosinolate metabolism and transport. *Plant Physiology and Biochemistry* **39**, 743-758.

Chen, L., DeVries, A. L. and Cheng, C. C. (1997) Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proceedings of the National Academy of Science* **94**, 3811-3816

Chen, S. and Halkier. B. A. (1999) Functional expression and characterization of the myrosinase MYR1 from *Brassica napus* in *Saccharomyces cerevisiae*. *Protein Expression and Purification* **17,** 414-420.

Chomczynski, P. and Sacchi, N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical Biochemistry* **162,** 156-159

Chrispeels, M. J. and Raikhel, N. V. (1991) Lectins, lectin genes, and their role in plant defense. *The Plant Cell* **3,** 1-9.

Chui, C-F and Falco, S. C. (1995) A new methionine rich seed storage protein from maize. *Plant Physiology* **107,** 291

Clark, J. M. (1988) Novel non-templated nucleotide addition reactions catalyzed by prokaryotic and eukaryotic DNA polymerases. *Nucleic Acids Research* **16**, 9677-9686

de Koning, A. P., Brinkman, F. S., Jones, S. J. and Keeling, P. J. (2000) Lateral gene transfer and metabolic adaptation in the human parasite *Trichomonas vaginalis*. *Molecular Biology and Evolution* **17**, 1769–1773.

Domon, B. and Aebersold, R. (2006) Mass spectrometry and protein analysis. *Science* **312,** 212-217

Dunwell, J. M. (1998) Cupins: a new superfamily of functionally diverse proteins that include germins and plant storage proteins. *Biotechnology and Genetic Engineering Reviews* **15**, 1-32

Dunwell, J. M., Khuri, S. and Gane., (2000) Microbial relatives of the seed storage proteins of higher plants: conservation of structure and diversification of function during evolution of the cupin superfamily. *Microbial and Molecular Biology Reviews* **64**, 153-179

Dunwell, J. M., Purvis, A. and Khuri, S. (2004) Cupins: the most functionally diverse protein superfamily? *Phytochemistry* **65**, 7-17

Eriksson. S., Elk, B., Xue, J., Rask, L. and Meijer, J. (2001) Identification and characterization of soluble and insoluble myrosinase isoenzymes in different organs of *Sinapis alba*. *Physiologia Plantarum* **111,** 353-364

Eriksson, S., Andreasson, E., Ekbom, B., Graner, G., Pontoppidan, B., Taipalensuu, J., Zhang, J., Rask, L. and Meijer, J. (2002) Complex formation of myrosinase isoenzymes in oilseed rape seeds are dependent on the presence of myrosinase-binding proteins. *Plant Physiology* **129**, 1592-1599

Fahey, J. W., Zalcman, A. T. and Talalay, P. (2001) The chemical diversity of glucosinolates and isothiocyanates. *Phytochemistry* **56**, 5-51.

Felsenstein, J. (2004) Inferring Phylogenies, Sinauer Associates, Inc., Sunderland, Massachusetts.

Felsenstein, J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

Foo, H. L., Grønning, L. M., Goodenough, L., Bones, A. M., Danielsen, B. E., Whiting, D. A. and Rossiter, J. T. (2000) Purification and characterization of epithiospecifier protein from *Brassica napus*: enzymatic intramolecular sulphur addition within alkenyl thiohydroximates derived from alkenyl glucosinolate hydrolysis. *FEBS Letters* **468**, 243-246.

Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M. R., Appel R. D. and Bairoch A. (2005) *Protein Identification and Analysis Tools on the ExPASy Server;* (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press (2005) 571-607

Gherardini, P. F., Wass, M. N., Helmer-Citterich, M. and Sternberg, M. J. E. (2007) Convergent evolution of enzyme active sites is not a rare phenomenon. *Journal of Molecular Biology* **372**, 817-845

Gilbert, W. (1978) Why genes in pieces? *Nature* **271**, 501

Gordon, J. A. (1972) Denaturation of globular proteins. Interaction of guadinium salts with three proteins. *Biochemistry* **11,** 1862-1870

Grishin, N. V. (2001) Fold change in evolution of protein structures. *Journal of Structural Biology* **134**, 167-185

Hanahan, D. (1983) Studies on transformation of *Escherichia coli* with plasmids. *Journal of Molecular Biology* **166**, 557-580

Henzel, W. J., Billeci, T. M, Stults, J. T., Wong, S. C., Grimley, C. and Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings of the National Academy of Science* **90**, 5011-5015

Hoekstra, H. E., Hirschmann, R. J., Bundey, R. A., Insel, P. A. and Crossland, J. P. (2006) A single amino acid mutation contributes to adaptive beach mouse colour pattern. *Science* **313**, 101-104

Holder, M. And Lewis, P. O. (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics* **4**, 275-284

Hughes, A. L. (1994) The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London* B: Biological Sciences **256,** 119-124

Hunt, D. F., Yates III, J. R., Shabanowitz, J., Winston, S. and Hauer, C. R. (1986) Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Science* **83**, 6233-6237

Husebye, H., Arzt, S., W.P. Burmeister, W. P., Haertel, F. V., Brandt, A., Rossiter, J. T. and Bones, A. M. (2005) Crystal structure at 1.1. ANG. resolution of an insect myrosinase from *Brevicoryne brassicae* shows its close relationship to b-glucosidases, *Insect Biochemistry and Molecular Biology* **35**, 1311–1320.

Iori, R., Rollin, P., Streicher, H., Thiem, J. and Palmieri, S. (1996) The myrosinase-glucosinolate interaction mechanism studied using some synthetic competitive inhibitors. *FEBS Letters* **385,** 87-90

Jander, G., Cui, J., Nhan, B., Pierce, N. E., and Ausubel, F. M. (2001) The *TASTY* locus on chromosome 1 of *Arabidopsis* affects feeding of the insect herbivore *Trichoplusia ni. Plant Physiology* **126,** 890-898

James, D. C. and Rossiter, J. T. (1991) Development and characteristics of myrosinase in *Brassica napus* during early seedling growth. *Physiologia Plantarum* **82,** 163-170

Jones, A. M. E., Bridges, M., Bones, A. M., Cole, R. And Rossiter, J. T. (2001) Purification and characterization of a non-plant myrosinase from the cabbage aphid Brevicoryne brassicae (L.). *Insect Biochemistry and Molecular Biology* **31**, 1–5

Jones, A. M. E., Winge, P., Bones, A. M., Cole, R. and Rossiter, J. T. (2002) Characterization and evolution of a myrosinase from the cabbage aphid *Brevicoryne brassicae*. *Insect Biochemistry and Molecular Biology*. **32,** 275-84.

Jorgensen, L. B., Behnke, H. –D. and Mabry, T. J. (1977) Protein-accumulating cells and dilated cisternae of the endoplasmic reticulum in three glucosinolate-containing genera: *Armoracia, Capparis, Drypetes*. *Planta* **137**, 215-224

Kazana, E., Pope, T. W., Tibbles, L., Bridges, M., Pickett, J. A., Bones, A. M., Powell, G and Rossiter, J. T. (2007) The cabbage aphid: a walking mustard oil bomb. *Proceedings of the Royal Society B* **274**, 2271-2277

Keck, A. S. and Finley, J. W. (2004) Cruciferous vegetables: cancer protective mechanisms of glucosinolate hydrolysis products and selenium. *Investigative Cancer Therapies* **3,** 5-12

Kelly, P. J., Bones, A. M. and Rossiter, J. T. (1998) Subcellular immunolocalisation of the glucosinolate sinigrin in seedlings of *Brassica juncea. Planta* **206**, 370-377

Kinch, L. N. and Grishin, N. V. (2002) Evolution of protein structure and functions. *Current Opinions in Structural Biology* **12**, 400-408

Kjaer, A., and Friis, P. (1962) Isothiocyanates XLIII. Isothiocyanates from *Putranjiva roxburghii* Wall. including (S)-2-methylbutyl isothiocyanate, a new mustard oil of natural derivation. *Acta Chemica Scandinavica* **16**, 936-946

Kliebenstein, D. J., Kroyman, J. and Mitchell-Olds, **T.** (2005) The glucosinolate-myrosinase system in an ecological and evolutionary context. *Current Opinion in Plant Biology* **8,** 264-271.

Kolkman, J.A. and Stemmer, P.C. (2001) Directed evolution of proteins by exon shuffling. *Nature Biotechnology* **19**, 423-428

Krishnan, H. B., Jiang, G., Krishnan, A. H. And Wiebold, W. J. (2000) Seed storage protein composition of non-nodulating soybean (*Glycine max* (L.) Merr.) and its influence on protein quality. *Plant Science* **157,** 191-199

Kwok, S., Chang, S-Y., Sninsky, J. J. and Wang, A. (1994) A guide to the design and use of mismatched and degenerate primers. *PCR Methods and Applications* **3**, 39-47

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680-685

Lambrix V., Reichelt M., Mitchell-Olds T., Kliebenstein D. J., Gershenzon J. (2001) The *Arabidopsis* epithiospecifier protein promotes the hydrolysis of glucosinolates to nitriles and influences *Trichoplusia ni* herbivory. *Plant Cell* **13**, 2793-2807

Larkin, M. A., Black shields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. and Higgins, D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947-2948

Lazzeri, L., Curto, G., Leoni, O. and Dallavalle, E. (2004) Effects of Glucosinolates and Their Enzymatic Hydrolysis Products via Myrosinase on the Root-knot Nematode Meloidogyne incognita (Kofoid et White) Chitw. Journal of Agricultural and Food Chemistry **52**, 6703-6707

Leander, B. (1998) Different modes of convergent evolution reflect phylogenetic distances: a reply to Arendt and Reznick. *Trends in Ecology and Evolution* **23,** 481-482

Lenman, M., Rödin, J., Josefsson, L. G. and Rask, L. (1990) Immunological characterization of rapeseed myrosinase. *European Journal of Biochemistry* **194,** 747-753.

Li, X. and Kushad, M. M. (2005) Purification and characterization of myrosinase from horseradish (*Armoracia rusticana*) roots. *Plant Physiology and Biochemistry* **43,** 503-511.

Linhart, C. and Shamir, R. (2005) The degenerate primer design problem : theory and application. *Journal of Computational Biology* **12,** 431-456

Long, M. (2001) Evolution of novel genes. *Current Opinion in Genetics and Development* **11,** 673-680

Long, M. and Langley, C. H. (1993). Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* **260**, 91–95

Long, M., Betrán, E., Thornton, K. and Wang, W. (2003) The Origin of New Genes: Glimpses from the young and old. *Nature Review Genetics* **4**, 865-879

Lüthy, J. and Benn, M. H. (1977) Thiocyanate formation from glucosinolates: a study of the autolysis of allyglucosinolate in *Thlaspi arvense* L. seed flour extract. *Canadian Journal of Biochemistry* **55,** 1028-1031

MacGibbon D. A. and Allison, **R. M.** (1970). A method for the separation and detection of plant glucosinolases (myrosinases). *Phytochemistry* **9**, 541-544

Marchler-Bauer, A., Anderson, J. B., DeWeese_scott, C., Fedorova, N. D., Geer, L. Y., He, S., Hurwitz, D. I., Jackson, J. D., Jacobs, A. R., Lancxycki, C. J., Liebert, C. A., Liu, C., Madej, T., Marchler, G. H., Mazumder, R., Nikolskaya, A. N., Panchemko, A. R., Rao, B. S., Shoemaker, B. A., Simonyan, V., Song, J. S., Thiessen, P. A., Vasudevan, S., Wang, Y., Yamashita, R. A., Yin, J. J and Bryant, S. H. (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acid Research* **31**, 383-387

McLellan, T. (1982) Electrophoresis buffers for polyacrylamide gels at various pH. *Analytical Biochemistry* **126**, 94-99

Meinke, D. W., Chen, J. and Beachy, R. N. (1981) Expression of storage-protein genes during soybean development. *Planta* **153,** 130-139

Meulenbeld, G. H. and Hartmans, S. (2001) Thioglucosidase activity from *Sphingobacterium* sp. strain OTG1. *Applied Microbiology and Biotechnology* **56**, 700-706

Mithen, R. (2001) Glucosinolates – biochemistry, genetics and biological activity. *Plant Growth Regulation* **34,** 91-103

Moran, J. V., DeBerardinis, R. J. and Kazazian Jr., H. H. (1999) Exon shuffling by L1 retrotransposition. *Science* **283**, 1530-1534

Nurminsky, D. I., Nurminskaya, M. V., De Aguiar, D. and Hartl, D. L. (1998). Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* **396**, 572–575

Ohno, S. (1973) Ancient linkage groups and frozen accidents. *Nature* **244,** 259- 262

Page, R. D. M. and Holmes, E. C. (1998) Molecular evolution: a phylogenetic approach, Wiley-Blackwell

Perrière, G. and Gouy, M. (1996) WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie*, **78**, 364-369

Pontoppidan, B., Ekbom, B., Eriksson, S. and Meijer, J. (2001) Purification and characterisation of myrosinase from the cabbage aphid *Brevicoryne brassicae*, a brassica herbivore. *European Journal of Biochemistry* **268,** 1041-1048.

Pratt, C., Pope, T. W., Powell, G., Rossiter, J. T. (2008) Accumulation of Glucosinolates by the cabbage aphid *Brevicoryne brassicae* as a defence against two coccinellid species. *Journal of Chemistry and Ecology* **34**, 323-329

Rakariyatham, N. and Sakorn, P. (2002) Biodegradation of glucosinolates in brown mustard seed meal (*Brassica juncea*) by *Aspergillus* sp. NR-4201 in liquid and solid state cultures. *Biodegradation* **13**, 395-399

Rask L, Andreason, E., Ekbom, B., Eriksson, S., Pontoppidan, B. and Meijer, J. (2000) Myrosinase: gene family evolution and herbivore defense in Brassicaceae. *Plant Molecular Biology* **42,** 93-113.

Ratzka, A., Vogel, H., Kliebenstein, D. J., Mitchell-Olds, T. and Kroymann, J. (2002) Disarming the mustard oil bomb. *Proceedings of the National Academy of Science* **99**, 11223-11228

Reid, G. E. and McLuckey, S. A. (2002) 'Top down' protein characterization via tandem mass spectrometry. *Journal of Mass Spectrometry* **37**, 663-675

Rossolini, G. M., Cresti, S., Ingianni, A., Cattani, P., Riccio, M. L. and Satta, G. (1994) Use of deoxyinosine-containing primers for polymerase chain reaction based on ambiguous sequence information. *Molecular and Cellular Probes* **8**, 91-98

Samonte, R. V. and Eichler, E. E. (2002) Segmental duplications and the evolution of the primate genome. *Nature Reviews Genetics* **3**, 65-72

Saitou, N. and Nei, M. (1989) The neighbour-joining method: a new method for reconstructing trees. *Molecular Biology and Evolution* **4**, 406-425

Schluter, D., Clifford, E. A., Nemethy, M. And McKinnon, J. S. (2004) Parallel evolution and inheritance of quantitative traits. *The American Naturalist* **163**, 809-822

Shrewy, P. R., Napier, J. A. and Tatham, A. S. (1995) Seed storage proteins: structure and biosynthesis. *The Plant Cell* **7,** 945-956

Sigrist, C. J. A., Cerutti, L., de Castro, E.,Langendijk-Genevaux, P. S., Bullard, V., Bairoch, A. and Hulo, N. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acid Research* **38,** 161D-166D

Snel, B., Bork, P. and Huynen, M. (2000) Genome evolution: gene fusion versus gene fission. *Trends in Genetics* **16**, 9-11

Steel, M. and Penny, D. (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and Evolution* **17**, 839-850

Stevens, P. F. (2001) Angiosperm phylogeny website: http://www.mobot.org/MOBOT/research/APweb/, Version 9, 2008

Stewart, C-B (1993) The powers and pitfalls of parsimony. *Nature* **361**, 603-607

Sudhof, T. C., Goldstein, J. L., Brown, M. S., Russell, D. W. (1985) The LDL receptor gene: a mosaic of exons shared with different proteins. *Science* **228**, 815-822

Taipalensuu, J., Andreasson, E., Eriksson, S. and Rask, L. (1997) Regulation of the wound-induced myrosinase-associated protein transcript in *Brassica napus* plants. *European Journal of Biochemistry*. **247**, 963-971.

Taipalensuu, J., Eriksson, S. and Rask, L. (1997) The myrosinase-binding protein from *Brassica napus* seeds possesses lectin activity and has a highly similar vegetatively expressed wound inducible counterpart. *European Journal of Biochemistry*. **250,** 680-688.

Thangstad, O. P., Gilde, B., Chadchawan, S., Seem, M., Husebye, H., Bradley, D. and Bones, A. M. (2004) Cell specific, cross-species expression of myrosinse in *Brassica napus*, *Arabidopsis thaliana* and *Nicotiana tabacum*. *Plant Molecular Biology* **54,** 597-611.

Todd, A. E., Orengo, C. A. and Thornton, M. (1999) Evolution of protein function, from a structural perspective. *Current Opinions in Chemical Biology* **3**, 548-556

van Rijk, A. A. F., de Jong, W. W. and Bloemendal, H. (1999) Exon shuffling mimicked in cell culture. *Proceedings of the National Academy of Science* **96**, 8074-8079

van Rijk, A. and Bloemendal, H. (2003) Molecular mechanisms of exon shuffling: illegitimate recombination. *Genetica* **118**, 245-249

Wang, W., Yu, H. and Long, M (2004) Duplication-degeneration as mechanism of gene fission and the origin of new genes in *Drosophila* species. *Nature Genetics* **36**, 523-527

Watkins, N. E., Jr. and SantaLucia, J., Jr. (2005) Nearest-neighbour thermodynamics of deoxyinosine pairs in DNA duplexes. *Nucleic Acids Research* **33**, 6258-6267

Wittstock, U., Agerbirk, N., Stauber, E. J., Olsen, C. E., Hippler, M., Mitchell-Olds, T., Gershenzon, J. and Vogel, H. (2004) Successful herbivore attack due to

metabolic diversion of plant chemical defense. *Proceedings of the National Academy of Science* **101**, 4859-4864

Wittstock, U. and Burow, M. (2007) Tipping the scales – specifier proteins in glucosinolate hydrolysis. *IUBMB Life* **59,** 744751

Wittstock, U. and Halkier, B. A. (2002) Glucosinolate research in the *Arabidopsis* era. *Trends in Plant Science* **7**, 263-270

Wu, G., Fiser, A., Kuile, B. T., Sali, A. and Muller, M. (1999) Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenas. *Proceedings of the National Academy of Science* **96**, 6285-6290

Xue, J., Jorgensen, M., Pihlgren, U. and Rask, L. (1995) The myrosinase gene family in *Arabidopsis thaliana*: gene organization, expression and evolution. *Plant Molecular Biology* **27,** 911-922

Zhang , J. (2003) Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, **18,** 292-298

Zhang, J., Pontoppidan, B., Xue, J., Rask, L. and Meijer. J. (2002) The third myrosinase gene TGG3 in *Arabidopsis thaliana* is a pseudogene specifically expressed in stamen and petal. *Physiologia Plantarum* **115,** 25-34.

Zhu, J. and Schiestl, R. H. (1996) Topoisomerase I involvement in illegitimate recombination in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **16**, 1805-1812