National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

The Evolution of Angiosperm Actin Genes

by

Mário Moniz de Sá

Thesis submitted to the
School of Graduate Studies and Research
University of Ottawa
in partial fulfillment for the
Ph.D. degree in the
Ottawa-Carleton Institute of Biology

Ottawa, Ontario, Canada

September 28, 1995

ISBN 0-612-07854-X

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

In memory of my father

and grandfather

Acknowledgements

I would like to thank Guy Drouin for giving me the opportunity of doing a doctoral degree under his supervision, for his guidance in developing this thesis research, for the financial assistance that he provided over the course of 4 years, and in particular his patience. Also, I would like to thank the members of my supervisory committee, François Chapleau, Donal Hickey, and George Carmody, for their time, interest and useful discussions. As well, I would like to thank David Currie and Paul Clarck for statistical advice.

I am greatly endebted to the Biology Department, at the time headed by Dr. Tom Moon, for accepting me into the doctoral program, and providing me with financial assistance.

I also would like to thank the graduate students, Post Docs, and undergraduate students working in our lab. They include: Michael Ell, Jamie Nickerson, Claude "of the stream" Desruisseaux, Jean Rousseau, Pascal Mayer, Anick DeMoors, and Ghislaine Allard. They made working in the lab a lot of fun. I also would like to thank all my friends in the Biology Department.

Finally, I would like to thank my wife Claudia for her love, patience, understanding, and financial support during the course of my work. I am forever greatful to her for interrupting her busy career and life in Vancouver to come to the coldest capital city in the world.

Table of Contents

Abstract

Forty four actin genes from five angiosperm species, whose evolutionary relationships are well characterized, were PCR-cloned and sequenced. Phylogenetic analysis of 34 of these actin genes, along with those previously published, indicate that plant actin genes are monophyletic and underwent a rapid radiation early in land plant evolution. Six sets of putative orthologues have been identified and and their sequences were used to calculate rates of evolution. The synonymous rate of substitution (5.44 x 10^{-9}/site/year) is similar to that of other nuclear protein-encoding genes but the non-synonymous rate (0.13 x 10^{-9}/site/year) is 4-10 times higher than that of vertebrate actin genes. Relative-rate tests do not support a faster rate in plants than in vertebrates. Evidence is also provided that some members of the actin multigene family in maize are undergoing gene conversion. Finally, we show that some plant actin genes have undergone intron loss probably as a consequence of a gene conversion event between the genomic copy and the reverse transcript.

Résumé

Quarante-quatre gènes d'actine ont été clonés et séquencés. Ces gènes proviennent de cinq espèces d'angiosperme dont les relations évolutives sont bien établies. Une analyse phylogénique de ces 34 gènes, ainsi que des séquences déja publiées, montre que les gènes d'actine des angiospermes sont monophylétiques et que plusieurs d'entres eux furent dupliqués avant l'apparition des angiospermes. Leur taux de substitutions synonymes (5.44 x 10^{-9}/site/année) est semblable au taux des gènes nucléaires des espèces animales, mais leur taux de substitutions non-synonymes (0.13 x 10^{-9}/site/année) est 4-10 fois plus èlevè que pour les gènes d'actine des espèces animales. Des tests de taux relatifs ne confirment pas un taux évolutif plus rapide pour les angiospermes, comparativement au vertébrés. Nous présentons des résultats montrant que certains gènes de la famille multigénique du maïs sont sujets à la conversion génique. De plus, certains gènes d'actine dans les plantes ont perdu leurs introns. Ceci est probablement dû à une conversion génique entre la copie génomique de ces gènes avec une copie de leur ADN complémentaire.

List of Figures

List of Tables

Introduction

Actins, as the major component of the eukaryotic cytoskeleton, have played an important role in the evolution of eukaryotes. Many of the most basic features which are uniquely associated with eukaryotes, such as phagocytosis, mitosis and meiosis, cell fusion, and many other fundamental cellular processes, would not be possible without the involvement of actin. An understanding of the way the actin multigene family has evolved will undoubtedly contribute to our general understanding of eukaryotic evolution. Much is already known about the evolution of actins in vertebrates, invertebrates, insects, and other representatives of major lineages. In plants however, very few representatives of the actin multigene family have been cloned and sequenced and therefore our understanding of actin evolution in land plants is still lacking. In fact, our general knowledge about the evolution of multigene families in plants is still poorly understood, most of that knowledge comes from studies of animal genes.

This study focuses on the evolution of the actin multigene family in a group of taxonomically well defined angiosperms. By comparing the evolution of plant actin genes with those of other eukaryotes, as well as other multigene families, we hope to learn about aspects of actin gene evolution which may be either peculiar to plant actins or to multigene families in plants in general.

Evolution of genes and multigene families

Over the last 25 years our insight into the origin and nature of genes has increased dramatically and a paradigm has emerged which states that genes arise from preexisting genes. Genes do not just evolve spontaneously from a random sequence of nucleotides, but originate and evolve from shuffling and duplications of preexisting

genes. Over time, a constant rain of mutation may lead to sequence and functional divergence between duplicated genes which may themselves be duplicated again. As a consequence of this repeated process of gene duplication followed by subsequent differentiation, it is now believed that most nuclear proteins are encoded by multigene families which may themselves be part of supergene families.

Modern-day genes are thought to have evolved from smaller modular units which could recombine to form many different types of genes with different combinations of modular units. This is referred to as the exon theory of genes (Gilbert, 1987). Under this theory, introns are supposed to have facilitated the recombination of these modular units, called exons. Support for this theory comes mainly from the fact that in some genes the exons, which are separated from one another by introns, encode the functional domains of proteins, and in some cases, similar domains are found associated with other genes but in different combinations (Gilbert, 1987). For those groups of organisms which do not possess introns, such as eubacteria and archaebacteria, it is hypothesized that during the course of their evolution, their introns were lost in the interest of having more economic genomes (Gilbert, 1987). However, the location of introns do not always separate exons encoding separate functional domains. This has been used as evidence against the exon theory of genes (Stoltzfus et al., 1994; Weber and Kabsch, 1994). Opponents of the exon theory of consider that prokaryotes did not lose their introns, rather they never had them in the first place. Instead, eukaryotes acquired introns during their evolution (Cavalier-Smith, 1991a). This controversy persists. The increasing knowledge about the different types of introns and their distributions and mode of splicing will undoubtedly contribute to this debate (Cavalier-Smith, 1991a; Palmer and Logsdon, 1991). It is agreed by all, however, that all life forms share many genes in common indicating that genes have been present and evolving since the beginning of life on this planet.

It is likely that the first life forms possessed much smaller and simpler genomes and that modern genomes have evolved in complexity in part due to gene and genome duplication (Gilbert, 1987; Li, 1983). Newly duplicated genes would be free to evolve new functions. In some cases, duplicated genes diverged so much from one another that little sequence similarity persists (Li, 1983). In other cases, especially when the duplications are recent, homologies can still be detected among them. A multigene family is defined as a group of genes related to one another by relatively recent duplications but which still maintain a similar function. A supergene family consists of a group of genes or a group of multigene families which are related to one another by relatively ancient duplications but which have greatly diverged from one another, and that may now have different functions but are still recognizable as being related to one another. There are no definitive criteria to decide whether a gene belongs to a multigene family or a supergene family since there are varying degrees of relatedness, but it has been suggested that genes sharing more than 50% similarity at the amino acid level should be referred to as being part of a multigene family, whereas genes showing less than 50% similarity at the amino acid level should be referred to as being part of a supergene family (Li and Graur, 1991).

The evolutionary fate of duplicated genes can be quite varied depending on how evolutionarily constrained they are. Some duplicated genes may diverge and acquire new functions by virtue of being expressed differentially or by acquiring a new biochemical function (Li and Graur, 1991; Li, 1983). This may be attributable to the relaxation of evolutionary constraints on the redundant second copy of the gene which is now free to evolve. Undoubtedly, this has been a very important process in the evolution of more complex organisms from simple ones. However, gene duplication can also serve to increase the number of copies of genes that are required to produce large quantities of a specific protein (Ohta, 1991). For example, the histone genes must

be expressed in large amounts during the S phase of the cell cycle. Other examples include the rRNA and tRNA genes.

Highly repetitive genes, such as histones and rRNA genes, are generally very similar in sequence. One factor responsible for this may be purifying selection because such genes have very specific functional or structural requirements. Another factor which may lead to a homogenization of genes is concerted evolution which occurs as a result of unequal crossing over and gene conversion (Arnheim, 1983; Li, 1983). A duplicated gene which is not under the influence of purifying selection can also become a pseudogene. This type of gene has become nonfunctional by deleterious mutations and are free to accumulate mutations with neutral effect on the organism. Consequently, pseudogenes evolve very rapidly (Li, 1983).

Conservation of gene families can be interpreted as indicating a functional role (selection) while divergence indicates absence of a functional role (drift) (Ohta, 1991). A contrasting view, called molecular drive, states that members of gene families become homogenized and fixed as a consequence of molecular turnover mechanisms in the genome (Dover, 1982; Dover and Tautz, 1986). It is likely that all three processes are involved to varying extents in the maintenance of different gene families.

Actin Function

Actin is a ubiquitous protein found in all eukaryotes. Actin monomers polymerize dynamically to form microfilaments which interact with a variety of actin-binding proteins to form an actin cytoskeleton (Seagull, 1989). The function of microfilaments has been well established in animals, yeast, and algae where they play a crucial role in a variety of fundamental cellular processes (reviewed in Sheterline and Sparrow, 1994, and Seagull, 1989).

In plants, the role of actin has not been thoroughly studied. However, using compounds which specifically inhibit polymerization or depolymerization, a variety of functions have been implicated for plant actins. These include roles in such fundamental processes as cell division and mitosis, chromosomal condensation, maintenance of cell shape (i.e. cytoskeleton), intracellular transport, cytoplasmic streaming, locomotion, chloroplast aggregation and rotation, growth of pollen tubes and root hairs, and it has been found associated with the endoplasmic reticulum (reviewed by Reece et al., 1992 and Meagher, 1991; Seagull, 1989). Based on amino acid sequence conservation between animal and plant actins, their similar solubility, ultrastructure, and electrophoretic mobility properties, it is thought that actins in plant cells function in much the same way as in animal cells (Reece et al., 1992). These functions are fundamental to a eukaryotic way of life, and it is likely that the evolution of actins has played a major role in the evolution of eukaryotes. In fact, the evolution of an internal cytoskeleton is thought to represent one of the most important innovations in the evolution of eukaryotes. It enabled cells to become larger and to increase in genetic complexity. It enabled phagocytosis and thus the acquisition by eukaryotic cells of endosymbionts, as well as it allowed for cell fusion and thus making "sex" possible (Cavalier-Smith, 1991b).

The diversity in the function of the actin proteins is probably modulated by interaction with other proteins (Sheterline and Sparrow 1994; Seagull, 1989). Actin is known to interact with many proteins which modulate actin polymerization, anchoring, and motor function. Amino acid sequences thought to be involved in interacting with actin-binding proteins are conserved between vertebrate and plant actins, again implying similarity of function. These actin-binding proteins are thought to constitute strong selective constraints against change and explain the high degree of conservation

of actin genes across broad phylogenetic lines (Sheterline and Sparrow, 1994; Beifuss and Durica, 1992).

Actin genes in eukaryotes

Actin is encoded by a multigene family in all eukaryotes studied so far with the exception of some fungi (e.g. *Saccharomyces cerevisiae* and *Aspergillus nidulans*) and some protists (e.g. *Tetrahymena thermophila* and *Giardia lamblia*). In vertebrates, actins are divided into distinct isotypes broadly classified as cytoplasmic (cytoskeletal) and muscle actins (Kovacs and Zimmer, 1993; Reece et al., 1992). Cytoplasmic actins are predominantly found in non-muscle cells with at least two isotypes found in mammals (referred to as β and γ) and three or more in avians and amphibians. Muscle actins are classified as striated or smooth based upon the muscle type in which they predominate. There are two striated actins; the α-skeletal muscle actin which is mainly expressed in skeletal muscle tissue and the α-cardiac muscle actin which predominates in cardiac muscle. Two forms of actin are expressed in smooth muscle tissues; α-smooth muscle isoform predominates in vascular tissue and γ-smooth muscle isoforms predominating in smooth muscle cells of the genital and gastrointestinal tracts (Kovacs and Zimmer, 1993; Reece et al., 1992). Thus, in vertebrates the different members of the actin multigene family are differentially regulated. In addition to these 6 functional genes, in many animal genomes there are also 20-40 pseudogenes for cytoplasmic actin, many of which are presumed to be processed pseudogenes (Meagher, 1991).

All actin genes from invertebrates, regardless of function, more closely resemble vertebrate cytoplasmic actins. Even those actin genes expressed in the muscle

tissues of *C. elegans*, *Drosophila*, and sea urchin display a greater similarity to vertebrate cytoplasmic isoforms than they do to vertebrate muscle actins. In fact, muscle actins in *Drosophila* and *C. elegans* evolved independently of those in vertebrates (Kovilur et al., 1993; Vandekerckhove and Weber, 1984). This suggests that vertebrate muscle actins developed after the divergence of vertebrates and invertebrates (Reece et al., 1992; Cross et al., 1988). Fungal and protist actins are also more similar to vertebrate cytoplasmic actins (Reece et al., 1992). However, these cytoplasmic actins in invertebrates perform the same function as the vertebrate muscle actins. This suggests that the actin genes functioning to produce muscle proteins do so because they belong to a muscle ontogenic regulatory program, rather than because the gene coding sequence specifies a muscle or a cytoskeletal protein (Davidson et al., 1982). At present there is no evidence that the different actin isotypes have a different biochemical function (Sheterline and Sparrow, 1994). However, actin genes do show a pattern of tissue-specific and/or developmental stage-specific expression (Reece et al., 1992).

In flowering plants, actin is encoded by a large, diverse, and dispersed multigene family comprised of 8-40 genes (Thangavalu et al., 1993; Reece et al., 1992; Drouin and Dover, 1990; McLean et al., 1990; Nairn et al., 1988; Baird and Meagher, 1987: Bernatzky and Tanksley, 1986a,b). In *Petunia*, there is over 100 actin genes but most of them are thought to be nonfunctional (McLean et al., 1990). In tomato, *Arabidopsis*, potato, and *Petunia* actin genes have been localized to at least 5 linkage groups (McLean et al., 1990; Nairn et al., 1988; Bernatzky and Tanksley, 1986a).

As in other systems, the multiple copies of actin genes are thought to function in providing a more diverse pattern of gene regulation, rather than to direct a large amount of protein (Meagher, 1991; Hightower and Meagher, 1986). Based on the shared intron positions and the many short amino acid sequences unique to plant sequences, it is

believed that all the plant actin sequences are derived from a single ancestral sequence (Meagher, 1991). Relative to other nuclear genes, actin genes in general are highly conserved in both size and sequence (Reece et al., 1992; Meagher, 1991; Li and Graur, 1991). However, relative to animal actins, plant actin genes show a greater sequence divergence. This suggests that either they are an ancient multigene family or that they are evolving faster than those in animals. It has been suggested that plant actins are evolving at the same rate as those in fungi and animals and that duplication and divergence of actin genes began earlier in plants than in animals and fungi (Reece et al., 1992; Meagher, 1991).

The early diversification of the actin multigene family in plants is thought to have occurred concomitant with the early diversification of tissues, cell types, and development of the earliest plants (Meagher, 1991). As new tissues arose the requirements for actin function or patterns of expression changed. Gene duplication followed by subsequent mutation, and acted upon by drift and selection over time resulted in novel gene types. Based on the rates of expression of actin genes and the total amounts of actin in plant cell types, it is not likely that the role of the plant actin multigene family is to supply ample gene product to satisfy the needs of the cell (Meagher, 1991). Instead, the different members of the plant actin multigene family are differentially expressed to suit the needs of particular tissues and cell types (Thangavalu et al., 1993; Meagher, 1991; Hightower and Meagher, 1985). For example, in tobacco there is an actin gene which is temporally and spatially regulated such that it is only expressed in pollen (Thangavalu et al., 1993). The differential expression of actin transcripts has also been observed in soybean, (McLean et al., 1990; Hightower and Meagher, 1985), rice (McElroy et al., 1990), and witchweed (Wolf and Timko, 1994).

Origin of Actins

Actin sequences in eukaryotes are highly conserved in both their size and sequence. However, divergent actin-like sequences (more than 50% divergent) have been identified from fungi, slime molds, invertebrates, vertebrates, and in a prokaryotic cyanobacter, *Spirulina platensis* (Sheterline and Sparrow, 1994; Schwob and Martin, 1992). Some of these divergent actins are essential for cell survival, and in vertebrates some have been associated with the microtubule cytoskeleton. The presence of these divergent actin-related proteins suggest that conventional actin evolved from a gene duplication event(s) very early in the evolution of eukaryotes.

Studies of the three-dimensional molecular structure of actin have shown that it shares structural similarities with a large group of other proteins (such as hexokinase and the 70kD-ATPase heat-shock protein), all having the common property of binding and hydrolyzing ATP (Sheterline and Sparrow, 1994; Flaherty et al., 1991; Holmes et al., 1990; Kabsch et al., 1990). Since prokaryotes also express a range of such proteins, actins probably represent just one of many descendants of a common ancestral gene encoding an ATP-binding protein very early on in cellular evolution.

Intron position in actin genes

The structure of actin genes is quite variable with respect to introns. The current catalogue of intron positions for over 100 actin genes consists of 40 individual positions (Weber and Kabsch, 1994). The total number of introns per gene varies from zero to eight. Some lower eukaryotes such *as Entamoeba histolyca, Tetrahymena thermophila*, two *Phytophthora* species, *Dictyostelium discoideum*, *Trypanosoma brucei*, and some of the actin genes from *Bombyx mori* and *Drosoplila melanogaster* contain no introns at all. In some cases only one intron is present, such as in the actin

gene from *Saccharomyces cerevisiae*. Still others, such as the actin gene from *Volvox carteri* and the vertebrate smooth muscle actin, have eight and seven introns, respectively.

Conservation in intron positions along the coding region of actin genes differs between different phylogenetic groups. In vertebrates the six actin genes are grouped into three pairs of homologues. The intron positions are conserved between homologous, but not between nonhomologous members. In addition, three of the seven intron positions known in vertebrates are conserved between all vertebrate actins. One of these conserved intron positions is also observed in a nematode (*Onchocerca volvulus*) and a platyhelminth (*Taenia solium*). For many species, such as *Drosophila melanogaster*, sea urchin, and *Caenorhabditis elegans* for which more than one actin gene is known, some variation is found in the position of introns. In animals as a whole there are 17 intron positions and none of these intron positions is 100% conserved among all members. Thus many intron loss or gain events have occurred in animals.

Actin gene structure in angiosperms

The typical structure of an angiosperm actin gene is illustrated in Figure 1. All functional plant actin genes characterized so far have only three introns located in identical positions within the coding region (Reece et al., 1992; Drouin and Dover, 1990). This suggests that all angiosperm actins known so far are derived from a single common ancestor which diverged after the animal-plant divergence (Reece et al., 1992; Meagher, 1991). The first intron position of angiosperm actin genes is shared with *Volvox carteri* and *C. elegans*. The second intron position is shared with two *Plasmodium* actin genes as well as a sea star actin gene and muscle actin genes in vertebrates. The last intron position is shared with both actin genes from the nematode,

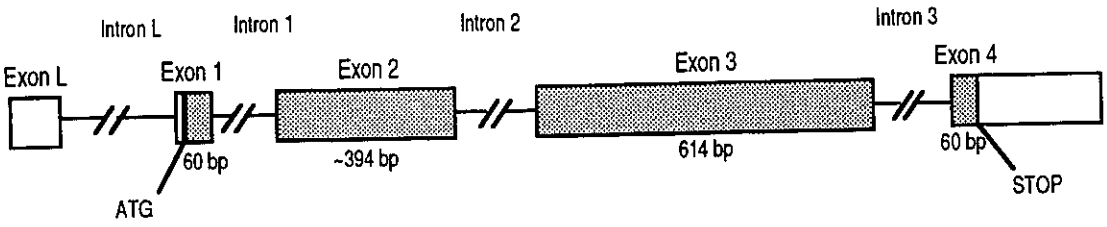

Figure 1: Structure of typical actin gene. The translated portions of exons are indicated by shaded boxes, untranslated portions by open boxes and introns are denoted by connecting lines. The approximate sizes of exons are also indicated. Intron sizes are variable. Start codon (ATG) and STOP codons are also indicated. Exon L and Intron L stand for leader exon and leader intron, respectively.

Onchocerca volvulus. Why these three intron positions are so conserved in angiosperm actins but not in other lineages in not known.

The size and sequence of plant actin introns are not conserved. The coding region is made up of four exons whose size and sequence are highly conserved. Typically, the coding region encodes a protein 377(±1) amino acids long. No significant insertions or deletions distinguish plant actin sequences.

Some well-characterized soybean and rice actin genes also contain an untranslated leader sequence which is interrupted by a relatively long leader intron (Pearson and Meagher, 1990; McElroy et al., 1990). Since other functional plant actin sequences also contain a potential intron splice acceptor sequence, this feature is probably shared among all plant actin genes (Meagher, 1991). Vertebrate actin genes also contain a leader exon interrupted by a leader intron which has been shown to possess enhancer activity (Kawamoto et al., 1988). A possible regulatory function of the leader intron in plant actin genes has been implicated from studies with rice (McElroy et al., 1990) and soybean actin genes (Meagher, 1991).

Why study actin genes

As already mentioned, actin genes are present only in eukaryotes and encode proteins whose functions are fundamental to eukaryotic life. This suggests that actin has played an important role in shaping eukaryotic evolution. Thus, any knowledge about the evolution of actin genes is likely to have some significance in terms of understanding the evolution of specific actin functions as well understanding the way in which eukaryotes evolved and became such a successful lineage showing incredible multicellular complexity.

Most of our knowledge about the evolution of genes and multigene families in eukaryotes has come from studies with animal systems. By studying the actin

multigene family in plants and comparing their evolutionary patterns with those in animals we may learn about possible differences in the way multigene families evolve in plants versus nonplant systems. Since much is already known about actin function and evolution in vertebrates and insects there is already a strong body of knowledge to serve as a basis for comparison.

In more practical terms, plant actins have several features which make them suitable and convenient to study. The conserved nature of actin genes enables the use of degenerate PCR primers to amplify actin genes from a diverse group of land plants. If the sequence strategy involves the use of internal primers then it is also convenient and less costly to have fewer sets of sequencing primers which work on as many different templates as possible. Also, actin DNA sequences from widely divergent species are easy to align due to their length conservation.

Plant actin genes are highly conserved and evolve at a rate which may be suitable to investigate events which happened early in the evolution of land plants. By identifying orthologous members between the major representatives of land plants, an inference can be made from the gene trees about the relationships between the major land plants and perhaps even elucidate the phylogenetic origins of the angiosperms. In addition, by comparing the function of orthologous members it is possible to determine whether the same function is conserved between orthologous members. This sort of information is useful because, if function is indeed conserved between orthologues, it may serve as the basis for extrapolation and prediction on the possible function of particular orthologues in different lineages.

When functional genes become nonfunctional, that is they become pseudogenes, their pattern of evolution undergoes a change based on the fact that the selective constraints, which may have been very rigid, have been relaxed. The evolution of pseudogenes is not well characterized in plants. The fact that an actin

processed pseudogene has been isolated in potato, and that actin pseudogenes are thought to be common in other plant species makes it likely that plant actin pseudogenes will be identified and be useful in understanding particular aspects of multigene family evolution as well as the nature of nucleotide change in the absence of purifying selection in plants.

Objectives of this study

Very few actin genes have been cloned and completely sequenced from any one plant species. As a consequence, our understanding of the evolution of actin genes in plants is at best patchy and incomplete. Previously, no actin orthologues between relatively recently diverged species had been identified with which to determine accurately the rate of evolution of plant actin genes. This has undoubtedly been in part due to the relatively large size of the actin multigene family in plants.

The objectives of this study were to clone, using PCR, and sequence as many actin genes as possible from five angiosperm species whose phylogenetic relationships are relatively well understood (Figure 2). These include four dicot species, three of which belong to the family Solanaceae (*Solanum tuberosum, Lycopersicon esculentum, and Nicotiana tabacum*) and one of which belongs to the family Fabaceae (*Glycine max*). The fifth species is a monocot belonging to the family Poaceae (*Zea mays*). These particular species were also chosen for study in part because there were already a number of actin genes cloned from them which can be added to our data set.

Phylogenetic analysis of actin genes using both distance-based and parsimony methods have been used to gain some understanding of the pattern of actin gene duplication in plants, as well as to identify putative orthologous genes. These orthologues have been used to test for rate constancy and to determine the rates of

evolution. Rate constancy has also been tested at various taxonomic levels and the rates of evolution in plants has been compared to those in fungi and animals.

Concerted evolution, via gene conversion or unequal crossing-over events, is a widespread and well documented phenomena resulting in sequence homogenization between members of a multigene family (Arnheim, 1983). Gene conversion has been implicated in the evolution of sea urchin actin genes (Durica et al., 1988; Crain et al., 1987). Gene conversion events can lead to erroneous interpretations of phylogenetic trees. In addition, rates of evolution will be underestimated if calculated from genes which have been influenced by concerted evolution. Therefore, plant actin genes have been analyzed to determine if concerted evolution has been a factor in their evolution.

Finally, phylogenetic analysis of plant actin genes along with those of other eukaryotic lineages has provided a broader understanding of actin gene evolution in eukaryotes as well as providing some insights into eukaryotic evolution as inferred from actin gene trees.

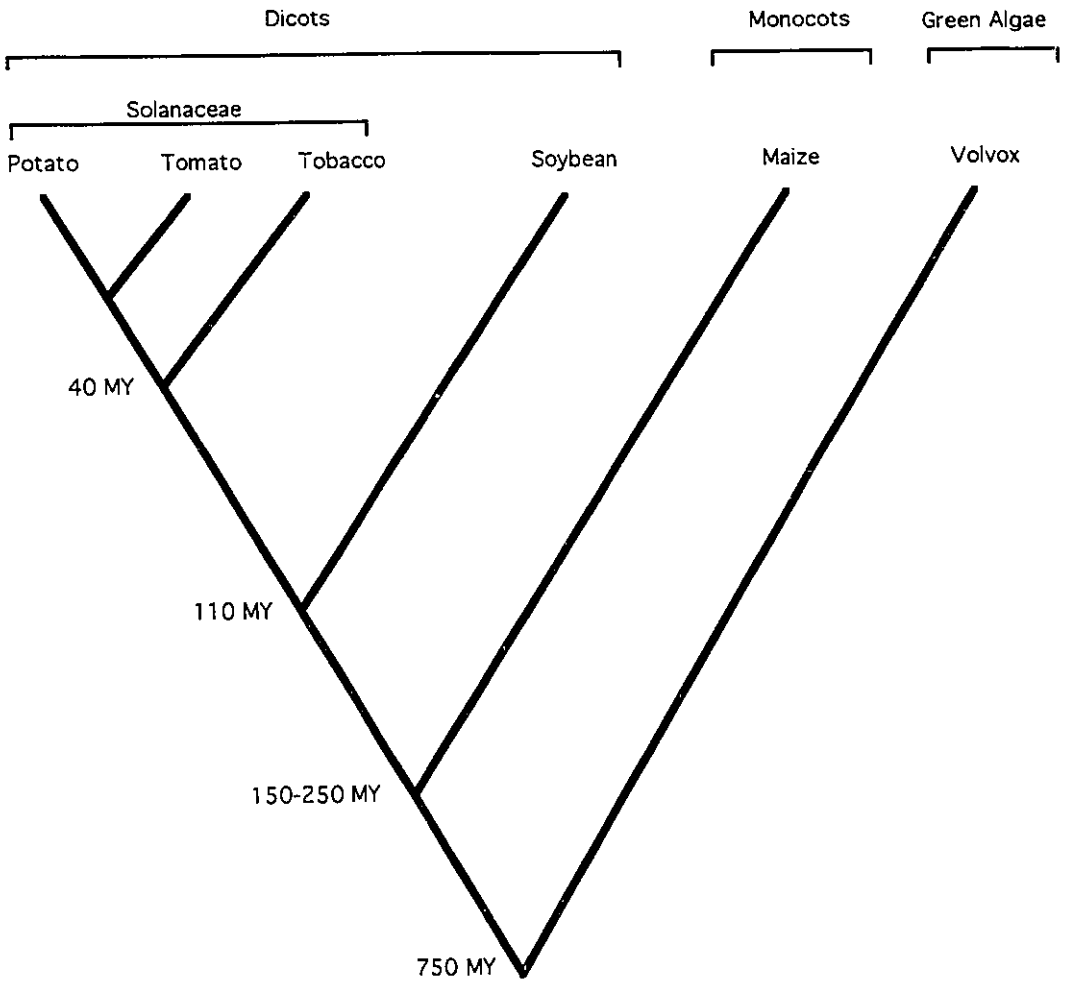

Figure 2: Phylogenetic tree of plant species used in this study showing approximate times of divergence (MY= millions of years). See materials and methods for details. Adapted from Meagher et al., (1989).

MATERIALS AND METHODS

Nucleic Acid Isolation

Genomic DNA was isolated from *Solanum tuberosum, Lycopersicon esculemtum, Nicotiana tabaccum, Glycine max* and *Zea mays* using a procedure adapted from Parsons et al. (1989). Plant tissues were ground in liquid nitrogen with a mortar and pestle to a fine powder and resuspended in extraction buffer (2 ml/gram of fresh weight of 100 mM Tris-HCl pH 8.0, 20 mM EDTA pH 8.0, 0.5 M NaCl, 0.5% sodium dodecyl sulphate (SDS), 0.5% 2-mercaptoethanol, and 1/5 vol of phenol). This extraction mixture was incubated at 65°C for 30 min and then centrifuged for 5 min at 8,000 rpm. The supernatant was extracted twice with 1:1 phenol:chloroform and extracted again 2-3 times with chloroform:isoamyl alcohol (24:1) with centrifugation for 10 min at 8,000 rpm following each extraction. At this point genomic DNA was precipitated with 1/3 volume of 10 M ammonium acetate (NH_3OAc) and 2 volumes of 95% ethanol and spooled out using a glass rod. The DNA was washed with 70% ethanol, resuspended in TE pH 8.0 containing 50 µg/ml of RNAse and incubated at 65°C for 1 hour. This solution was then extracted with phenol:chloroform and the DNA precipitated as described above before being resuspended in TE pH 8.0.

PCR Amplification

Actin sequences from potato, tomato, tobacco, soybean and maize were amplified by the polymerase chain reaction (PCR) using degenerate primers. The amino and carboxyl ends of plant actins are extremely well conserved. Based on this fact, degenerate primers were designed using the plant actin sequences already available. The sense primer (a 23 mer) hybridizes to the region from the first nucleotide of codon 12 through to the first two nucleotides of codon 19 (in the first exon). The antisense

primer (a 29 mer) hybridizes to the region from the first nucleotide of codon 357 through to the first two nucleotide positions of codon 366 (in the fourth exon). Note that only the sequences internal to exon 1 and exon 4 will be amplified and thus available for phylogenetic study.

PCR was carried out by placing approximately 150 ng of genomic DNA, 150 pmol each of the forward [5' TG(T/C) GA(T/C) AA(T/C) GGI ACI GGI ATG GT 3'] and reverse [5' TC(A/G) TC(A/G) TA(T/C) TCI IC(T/C) TTI GII ATC CAC AT 3'] degenerate primers, 8 µl of 2.5 mM deoxynucleotide triphosphate (dNTP) mixture, 4.5 µl of 50 mM $MgSO_4$, 10µl of 10X Taq PCR buffer (Promega), and the volume brought up to 100µl with water. The mixture was boiled for 5 minutes, quenched on ice for 5 minutes and centrifuged for 15 seconds before adding 0.5 units of Taq polymerase (Promega) and overlaying with 3 drops of parafin oil (BDH). PCR was carried for 30 cycles of 1 min at 94°C, 1 min at 50°C (or 2 min at 54°C), and 3 min at 72°C. A final cycle was then carried out for 15 min at 72°C to ensure complete extension. The samples were then extracted with 1 volume of 1:1 phenol: chloroform. One tenth volume of each PCR reaction was run on a 0.8% agarose gel with 1X TBE as a running buffer. The gel was stained with ethidium bromide and the amplified DNA visualized under UV.

Blunt-end cloning of PCR fragments

PCR fragments were made ready for cloning using the DOUBLE GENECLEAN protocol (BIO 101 Inc.). This procedure involves the purification of the PCR fragments using the GENECLEAN glass milk beads to remove excess nucleotides and primers left over from the PCR reaction. The ends of the PCR fragments were made blunt for blunt-end cloning using 25µl of PCR reaction eluate, 10 µl of 10X pol I buffer (0.5 M Tris pH 7.5, 0.1M $MgCl_2$, 10mM DTT, 0.5mg/ml BSA, and 200 µM

dNTPs), 10 µl of 10mM ATP, and 10 units each of T4 polynucleotide kinase (Pharmacia) and DNA polymerase I (Pharmacia). The final volume was adjusted to 100µl and incubated at 37°C for 1 hour. The reaction was stopped with 1µl of 0.5M EDTA pH8, GENECLEANed a second time, and resuspended in 25 µl of water.

The cloning vector used was the Bluescript SK^+/KS^+ phagemid (Stratagene). This vector contains many useful features which were exploited in this study. It contains an ampicillin resistance gene for antibiotic selection of the phagemid vector. It also contains the LacZ gene along with its promoter to provide α-complemetation for blue/white color selection of recombinant phagemids (see below). Bluescript also contains the f1 filamentous phage origin of replication allowing the recovery of single stranded DNA (ssDNA) when a host strain is co-infected with a helper phage (thus the name phagemid). The ssDNA was used for DNA sequencing. Finally, the Bluescript phagemid comes in two forms, the SK^+ and KS^+, which differ in that the multiple cloning site of SK^+ is reversed relative to the f1 origin of replication. Thus by cloning into both SK^+ and KS^+ versions of the phagemid it is possible to obtain ssDNA of both strands.

Blunt-ended PCR fragments were ligated into the *Sma I* multiple cloning site of the SK^+ Bluescript plasmid. Ligation was carried out in a 10 µl volume using 6.5 µl of DNA, 1 µl of 10X one-for-all buffer (Pharmacia), 0.5 µl of 10 mM ATP, 1 µl of 10 mM hexamine cobalt chloride, 1 µl of blunt-ended dephosphorylated SK^+ Bluescript plasmid DNA, and 8 units of T4 ligase (Pharmacia). The reaction was incubated overnight at 10°C. The ligase was inactivated by heating at 65°C for 5 min.

The SK^+ Bluescript plasmid was blunt-ended by digestion with *Sma I*. The reaction was carried out in a 100 µl volume containing 10 µg of plasmid DNA, 10 µl of one-for-all buffer (Pharmacia), and 50 units of *SmaI* (Pharmacia). The reaction was

stopped by heat inactivation at 65°C for 5 min. The reaction was extracted with 1 volume of 1:1 phenol:chloroform, precipitated with 1/10 vol of 3M sodium acetate and 2 volumes of cold 95% ethanol, and resuspended in 90 µl of water. Dephosphorylation of the *Sma I* digested plasmid DNA was carried out in a 100 µl volume containing 10 µg plasmid DNA, 10 µl of 10X calf intestine alkaline phosphatase buffer (CIP buffer) (Boehringer Mannheim), and 3 units of alkaline phosphatase (Boehringer Mannheim). The reaction was incubated at 37°C for 15 min, another 3 units alkaline phosphatase were added and reincubated this time at 56°C for 45 min. The phosphatase was denatured by the addition of 1.5 µl of 0.5 M EDTA pH 8 and heatshocked at 75°C for 10 min. The plasmid DNA was then extracted twice with 1 volume of 1:1 phenol:chloroform, precipitated, washed with 70% ethanol and resuspended in 40 µl of TE pH 8.0.

Transformation into *E.coli*

To select for plasmids containing the PCR fragments as inserts in the polycloning site, the ligation products were transformed into a DH11S strain of E. coli (Gibco BRL) competent cells. One microliter of ligated products were added to 100 µl of freshly thawed competent cells and kept on ice for 30 min. The sample was then heatshocked for 2 min at 42°C, followed by the addition of 1 ml of LB medium (10 g NaCl, 10 g Tryptone and 5 g of yeast extract per 1 liter of water, and autoclaved) and incubated for 1 hour at 37°C. Samples were centrifuged for 15 seconds and resuspended in 100 µl of LB, and plated on LB solid media (1.5% agarose) containing 50 µg/ml of ampicillin, 0.5 µM IPTG (Isopropylthiogalactoside) and 0.01% X-Gal (5-bromo-4-choro-3-indolyl-β-D-galactoside). IPTG is an analog of lactose which can not be metabolized but can act as an inducer of the β-galactosidase gene (LacZ) present in

the polycloning site of the Bluescript plasmid. X-gal is a chromogenic compound which is an analog of lactose and is a substrate of β-galactosidase but can not act as an inducer of the LacZ gene. Thus *E.coli* colonies carrying the plasmid with no inserts will appear blue because the coding region of the LacZ gene is uninterrupted and will be able to produce a functional β-galactosidase product capable of braking down the chromogenic agent X-gal. Conversely, colonies containing a plasmid with an insert will appear white because the coding region of the LacZ gene is interrupted. These white colonies were picked for further study.

E. coli DH11S competent cells were prepared as follows. A single colony was innoculated into 5 mls of LB medium and grown overnight at 37°C. This culture was then used to innoculate 400 mls of LB medium and grown at 37°C for about 2-3 hours until the OD_{590} was at 0.375. The culture was aliquoted into prechilled 50 ml tubes, kept on ice for 30 min., centrifuged for 5 min at 2500 rpm at 4°C, and the cells resuspended in 10 mls of sterile ice-cold $CaCl_2$ solution (60 mM $CaCl_2$, 15% glycerol, and 10 mM PIPES pH 7.0). The centrifugation and resuspension step was repeated once more and the resuspended cells kept on ice for 30 min. The cells were once more centrifuged, resuspended in 2 mls of $CaCl_2$ solution, alliquoted into 100 µl volumes in 1.5 ml eppendorf tubes, and flash frozen in a dry ice ethanol bath. Cells were thawed out for use when needed. The transformation efficiency of these cells were typically > $1x10^5$ colonies per 1 µg of Bluescript plasmid.

Screening for clones

White colonies were screened for the presence of plasmids containing inserts using the alkali mini-prep method (Maniatis et al., 1982). Colonies were picked from the solid LB media using sterile tooth picks and inoculated into 2.5 mls of LB

containing 50 µg/ml of ampicillin and grown overnight at 37°C. Cultures were transferred into 1.5 ml eppendorff tubes, centrifuged for 30 secs and the supernatant aspirated off. Bacterial pellets were resuspended in 100 µl of solution I (1% glucose, 25 mM Tris-HCl pH 8.0, and 10 mM EDTA pH 8.0) by vortexing. Cells were lysed with 200 µl of solution II (0.2 M NaOH, and 1% SDS), mixed by gentle inversion, and kept on ice for 10 min. Cell debris was precipitated with 150 µl of solution III (60 mls of 5 M potassium acetate, 11.5 mls of glacial acetic acid and 28.5 mls of water), mixed by inversion and kept on ice for 5 min before being centrifuged for 5 min at 12000 rpm. Supernatant was then extracted with 1:1 phenol:chloroform, centrifuged for 5 min at 12 000 rpm. The plasmid DNA was precipitated with 2 volumes of 95% ethanol, centrifuged at 12 000 rpm for 10 min, and the resulting pellet was washed with 70 % ethanol. The plasmid DNA was dissolved in 50 µl of TE pH 8.0 containing 20 µg/ml of RNAse.

To detect if plasmids contained inserts, plasmid DNA was double digested with the restriction enzymes, *EcoRI* and *BamHI*, and the fragments separated on a 0.8% agarose gel and the DNA bands visualized by staining the gel with ethedium bromide and exposure to UV light. Restriction digests were carried out at 37°C for 1 hour in a 20 µl volume containing 2 µl of plasmid DNA, 4 µl of 10X One-for-all buffer (Pharmacia) and 2.5 units each of *EcoRI* and *BamHI*. All plasmids containing inserts 1 kilobase or longer were kept for further analysis. Cultures containing the desired plasmids were frozen at -80°C by adding sterile glycerol to a concentration of 15%.

Preparation of single stranded DNA

Single stranded DNA (ssDNA) used for sequencing was prepared as follows. An overnight colony was inoculated into 1.5 mls of LB containing 10 mM $MgCl_2$ and 50 µg/ml ampicillin and grown on a gyratory shaker at 200 rpm at 37°C. After 2-3

hours of growth 5 µl of M13KO7 (Gibco BRL) helper phage was added and the volume brought up to 6 mls with the above mentioned LB media. Cultures were grown overnight using the above mentioned conditions. The cells were pelleted by centrifugation at 12 000 rpm for 15 min, and the viral particles in the supernatant allowed to precipitate on ice for 1 hour by adding NaCl and Polyethylene glycol (PEG) to a concentration of 0.5 M and 4 %, respectively. The viral particles were pelleted by centrifugation at 12 000 rpm for 15 min, the supernatant aspirated off, and the viral pellet was resuspented with 400 µl of TE pH 8.0 (containing 0.5 % SDS, 2.5 µg/ml of proteinase K. and 20 µg/ml of RNAse) and incubated at 65°C for 2-3 hours. The viral debris was precipitated by the addition of 5 M potassium acetate and cooling on ice for 10 min. This mixture was centrifuged at 12 000 rpm for 5 min and the supernatant extracted with one volume of 1:1 phenol:chlrorform. The ssDNA in the aqueous phase was precipitayed using 3 volumes of 95% ethanol and centrifuged at 12 000 rpm for 15 min. Excess salt was washed off from the ssDNA pellet with 70% ethanol. The ssDNA was resuspended in 30 µl of TE pH 8.0 and visualized by running 2 µl on a 0.8% agarose gel stained with ethidium bromide and exposed to UV.

DNA sequencing strategy

Over 100 PCR fragments were cloned from each of the five species. Rather than sequencing all of these clones to identify unique actin sequences, these clones were initially divided up into different size classes and 1-5 clones from each size class were sequenced. Based on a partial sequence of the 5' or 3' ends unique clones were identified, cloned into the KS^+ vector, and ssDNA prepared from both DNA strands for further sequencing. Single strand DNA templates for each PCR clone were sequenced according to the dideoxynucleotide termination method (Sanger et al., 1977) using the T7 primer and reaction conditions employing T7 DNA polymerase (Pharmacia) or

Sequenase vs 2.0 (United States Biologicals, Cleveland, OH). Other primers were used to sequence internal regions of the actin clones (Figure 3).

Phylogenetic relationships among species

The phylogenetic relationships among the five angiosperm species examined are shown in Figure 2. These relationships are well established from analysis of morphological traits and are based on the nomenclature of Cronquist (1981). Estimates of divergence times from a common ancestor are based on megafossil and/or fossil pollen data and were obtained from Meagher et al. (1989), and likely represent underestimates of the true divergence times. Although the phylogenetic relationship between the two tribes Cestreae (Tobacco) and Solaneae (potato and tomato) are well established (D'Arcy, 1979), their evolution from a common ancestor is less well understood. Therefore, the time of divergence of these two tribes from a common ancestor will be assumed to be that of the origin of the Solanaceae family, 40 MY ago. The divergence of the Fabaceae (containing soybean) and the Solanaceae from the subclass Rosidae will be assumed to be 110 MY ago, a time just after the radiation of the angiosperm subclasses, 140 MY ago. The divergence time between monocots and dicots is still controversial, but it is generally thought to be in the order of 150-250 MY (Wolfe et al., 1989). The oldest known green-algal fossils are dated at 750 MY and the divergence between green algae and land plants is thought to have occurred about 750 MY ago.

Sequence manipulation and alignments

DNA sequences were entered into an IBM compatible computer from the autoradiograms using the data entry program *DNA Parrot DP100-PC* version 2.3. DNA sequences were manipulated (i.e, introns and third base of codons removed) and

A

Degenerate Sequencing Primers:

ACC1: 5' (C/G)A(A/G) ATI ATG TTT GA(GA) AC 3' (17mer)

ACN1: 5' TCA AAC ATI AT(C/T) T(C/G)I GTC AT 3' (20mer)

ACC2: 5' AA(A/G) AI(T/C) (T/C)A(T/C) GA(A/G) ITI CCI GAT GG 3' (23mer)

ACN2: 5' TCI GGI AI(C/T) TC(A/G) (T/A)(A/G)(A/G) IT(C/T) TT(C/T) TC 3' (23mer)

ACN3: 5' CC(T/C) (T/C)C(A/G) ATC CA(G/A) ACA CT(A/G) TA 3' (20mer)

ACC5: 5' AA(C/T) TGG GA(C/T) GA(C/T) ATG GAG AA 3' (20mer)

ACN6: 5' CIA TIG TIA TIA C(T/C)T GIC CAT C 3' (22mer)

B

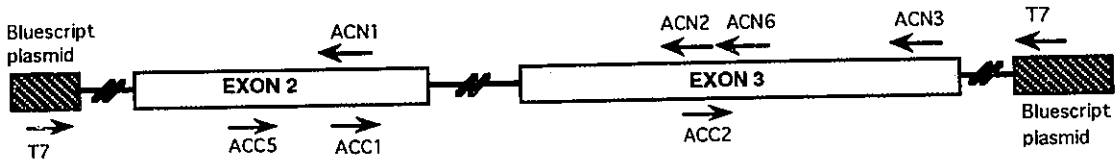

Figure 3: Sequencing strategy used to sequence both DNA strands of plant actin genes. (A) Sequence of degenerate primers and (B) their approximate homologous positions along the coding sequence. (Drawing not to scale).

coding sequences translated using *GDE*, a sequencing processing program running on a Unix workstation. Additional plant and other eukaryotic actin sequences were downloaded from Genbank and Embl data bases using the Entrez retrieval system and are listed in Tables 1, 2 and 3. Protein sequences were aligned manually on *GDE*. Manual alignment was not problematic because actin genes are highly conserved in size and sequence. The DNA sequences were aligned based on the protein alignment.

Sequence Analysis

Plant actin DNA sequences were analyzed for their total G+C content as well as the G+C content at each individual codon position and at their introns using the *CODONS* program (Lloyd and Sharp, 1992). The number of synonymous and non-synonymous substitutions for all pairwise comparisons were estimated according to the method of Li et al. (1985) using the program Li93 (Li, 1993a). The percent amino acid and nucleotide difference between actin genes, as well as the number of variable and parsimony informative sites were determined using the program *Molecular Evolutionary Genetics Analysis* Version 1.01 (*MEGA*) (Kumar et al., 1993). Analysis on intron homology was carried out using the SSEARCH program found in the FASTA program package (Pearson, 1994). Sequence analysis to detect gene conversion was kindly done by Frederik Pratt using a computer program which he is currently developing as part of his M.Sc. thesis.

Phylogenetic Analysis

Phylogenetic trees were built according to parsimony and neighbor-joining distance methods. Parsimony is a method of inferring phylogenies which finds the phylogeny on which the observed characters could have evolved with the least number of evolutionary changes (Felsenstein, 1988). In general, parsimony gives better

Sequence	Accession number	Organism
Vlvx	M33963	*Volvox carteri*
MAc1	J01238	*Zea mays*
SAc1	J01298	*Glycine max*
SAc3	J12097	*Glycine max*
Rac1	X15865	*Oryza sativa*
RAc2	X15864	*Oryza sativa*
RAc3	X15862	*Oryza sativa*
RAc7	X15863	*Oryza sativa*
DcAc1	S55573	*Daucus carota*
DcAc2	S55575	*Daucus carota*
SoAc1	X79378	*Sorghum vulgare*
PsAc1	X68649	*Pisum sativum*
PsAc2	X67666	*Pisum sativum*
AAc1	M20016	*Arabidopsis thaliana*
Tob1	X63603	*Nicotiana tobacum*
PoAc58	X55749	*Solanum tuberosum*
PoAc71	X55750	*Solanum tuberosum*
PoAc75	X55753	*Solanum tuberosum*
PoAc97	X55751	*Solanum tuberosum*
PoAc99	X00279	*Solanum tuberosum*
PoAc101	X55752	*Solanum tuberosum*

Table 1: Plant actins used in this study obtained from Genbank and Embl databases.

Sequence	Accession number
Giardia lamblia	L29032
Euplotes crassus	J04533
Oxytricha nova	M22480
Leishmania major	L16061
Trypanosoma brucei	M20310
Trypanosoma cruzi	U20234
Tetrahymena thermophila	M13939
Plasmodium falciparum 1	M19146
Plasmodium falciparum 2	M18817
Naegleria fowleri 1	M90311
Fucus disticus	U11697
Achlya bisexualis	X59936
Phytophthora megasperma	X15900
Entamoeba histolyca	M16339
Dictyostelium discoideum	X03281
Physarum polycephalum	M21500
Acanthamoeba castellanii	J01016
Trichoderma reesei	X75421
Aspergillus nidulans	M22869
Thermomyces lanuginosus	X07463
Candida albicans	X16377
Saccharomyces cereviciae	J01310
Kluyveromyces lactis	M25826
Filobasidiella neoformans	U10867
Schizosaccharomyces pombe	Y00447
Absidia glauca	M64729
Pneumocystis carinii	L21183
Caenorhabditis elegans	X16796
Drosophila melanogaster 2A	K00670
Drosophila melanogaster 5CX	K00667
Drosophila melanogaster 79B	M18829
Drosophila melanogaster 87E	X12452
Drosophila melanogaster 88F	M18830
Drosophila melanogaster 8F	M18826
Man cytoskeletal beta	M10277
Man cytoskeletal gamma	M24241
Man enteric gamma	X16940
Man skeletal A	J00068
Man vascularA	X13839
Xenopus laevis cardiac A	X03469
Xenopus laevis cytoskeletal 5	M24769
Xenopus laevis cytoskeletal 8	M24770
Xenopus laevis sarcomeric A2	X05392
Xenopus laevis sarcomeric A3	X12525
Strongylocentrotus purpuratus 2	V01350
Strongylocentrotus purpuratus IIb	M35323
Strongylocentrotus purpuratus IIIb	M35324
Aplysia cxalifornica	X52868
Cyprinus caprio beta	M24113
Hydra attenuata	M32364

Table 2: Protist, animal, and fungal actins obtained from Genbank/Embl databases used to generate phylogenetic tree shown in Figure16.

Sequence	Taxa	Accession number
Giardia lamblia	Protist 1	L29032
Euplotes crassus		J04533
Leishmania major		L16061
Trypanosoma brucei		M20310
Trypanosoma cruzi		U20234
Tetrahymena thermophila	Protist 2	M13939
Acanthamoeba castellanii		J01016
Entamoeba histolyca A		M16339
Entamoeba histolyca B		M16340
Entamoeba histolyca C		M16341
Entamoeba histolyca D		M19871
Naegleria fowleri 1		M90311
Naegleria fowleri 2		M90312
Plasmodium falciparum 1		M19146
Plasmodium falciparum 2		M18817
Fucus disticus		U11697
Phytophthora megasperma		X15900
Physarum polycephalum	Slime molds	X07792
Physarum polycephalum 35		M21500
Physarum polycephalum 5		M21501
Physarum polycephalum A10		M15272
Physarum polycephalum D		M59234
Physarum polycephalum B		X60788
Dictyostelium discoideum A12		X03282
Dictyostelium discoideum A3S1		X03283
Dictyostelium discoideum A3S2		X03284
Dictyostelium discoideum A8		X03281
Trichoderma reesei	Fungi	X75421
Candida albicans		X16377
Aspergillus nidulans		M22869
Saccharomyces cereviciae		J01310
Pneumocystis carinii		L21183
Kluyveromyces lactis		M25826
Schizosaccharomyces pombe		Y00447
Filobasidiella neoformans		U10867
Thermomyces lanuginosus		X07463
Caenorhabditis elegans 1	Animals	X16796
Caenorhabditis elegans 2		X16797
Caenorhabditis elegans 3		X16798
Drosophila melanogaster 2A		K00670
Drosophila melanogaster 5CX		K00667
Drosophila melanogaster 79B		M18829
Drosophila melanogaster 87E		X12452
Drosophila melanogaster 88F		M18830
Drosophila melanogaster 8F		M18826
Man beta		X00351
Man cytoplasmic beta		M10277
Man cytoskeletal gamma		M24241
Man cytoskeletal gamma 1		M19283
Man enteric smooth muscle gamma		X16940
Man adult skeletal alpha		J00068
Man vascular smooth muscle alpha		X13839
Xenopus laevis cardiac		X04669
Xenopus laevis cardiac alpha		X03469
Xenopus laevis cytoskeletal 5		M24769
Xenopus laevis cytoskeletal 8		M24770
Xenopus laevis sarcomeric alpha 2		X05392
Xenopus laevis sarcomeric alpha 3		X12525
Cyprinus caprio beta		M24113
Hydra attenuata		M32364

Table 3: Protist, animal, and fungal actins obtained from Genbank/Embl databases used to generate bar graph shown in Figure 21.

estimates of the true phylogeny when the rates of evolution are low and the different lineages are evolving at the same rate.

Neighbor-joining is a distance based method of inferring phylogenetic trees which sequentially identifies neighbor pairs that minimize the total length of the tree (Saitou and Nei, 1987). Because the neighbor-joining method does not choose the shortest tree from all or a large number of trees, but simply builds one by a stepwise clustering approach of neighbor pairs, it is an ideal method for dealing with relatively large data sets relatively quickly. In addition, the neighbor-joining method has been shown to be quite robust, especially when the rates of evolution differ between lineages (Saitou and Imanishi, 1989).

The aligned plant actin amino acid and nucleotide sequences were used to construct parsimony and distance-based phylogenetic trees. The aligned DNA data set (excluding third position of codons) was resampled using SEQBOOT to generate 100 data sets. Bootstrapping involves the creation of many artificial data sets, each the same size as the original one, by random sampling with replacement of the variable sites in the actual data sets. A recent study has shown that under a wide range of conditions, groups supported by bootstrap values of 70% have > 95% probability of representing true clades (Hillis and Bull, 1993). The program, DNAPARS, was used to calculate maximum parsimony trees from each of the 100 data sets and CONSENSE was used to obtain a majority-rule consensus tree using the *Volvox* actin gene as the outgroup.

The aligned plant DNA sequences were also used to calculate Neighbor-joining trees using the program MEGA (Kumar et al., 1993). The third position of codons was excluded from the analysis. Distances were estimated using Kimura's 2-parameter model. The output of this analysis was a majority-rule consensus tree based on 100 bootstrap data sets. Neighbor-joining distance trees were also calculated from the

number of non-synonymous substitutions for all pair-wise comparisons of plant actin genes. The non-synonymous substitutions were estimated using MEGA based on the Jukes-Cantor model. 100 bootstrap trees were calculated and a majority-rule consensus tree was obtained.

The aligned plant actin amino acid sequences were used to construct trees in the same way as the DNA sequences except that PROTPARS was used to calculate the maximum parsimony trees. The program MEGA (Kumar et al., 1993) was used to calculate neighbor-joining distance trees based on the proportion of amino acid differences for all pair-wise comparisons of plant actin genes. A majority-rule consensus tree was obtained from 100 bootstap data sets.

Aligned actin amino acid sequences from 45 eukaryotes were also used to generate a Neighbor-joining tree. However, because the size of this data set exceeded the memory limitations inherent in running MEGA on a PC, this analysis was carried out using PHYLIP (see below). Pairwise distances were calculated based on Dayhoffs PAM 250 matrix using the program PROTDIST. The Pam matrix is a weighted method of amino acid change which reflects the empirical probabilities of amino acid replacement for each possible change and is considered to be more biologically relevant than other mathematical models of amino acid change (Wilbur, 1985; Dayhoff et al., 1979). The distance matrix was then used to calculate the shortest Neighbor-joining tree using the NEIGHBOR program. Pairwise distances were also calculated (as above) for 100 bootstrap replicates and a majority rule consensus tree obtained. Because the consensus tree does not include information on branch lengths, bootstrap values over 50 from the consensus tree were transposed to branches of similar topology on the shortest Neighbor-joining tree.

The programs DNAPARS, PROTPARS, PROTDIST, NEIGHBOR, SEQBOOT, and CONSENSE are all part of the Phylogeny Inference Package, version 3.5 (PHYLIP) (Felsenstein, 1993).

Relative rate tests

The relative rate test is used to determine if sequences evolve at the same rate relative to a known outgroup. The relative rate test was done according to the method of Li and Tanimura (1987). The standard error of K_{13}-K_{23} was calculated as the square-root of the variance. The variance was calculated as follows: let,

$$K_{03} = \frac{\left(K_{13} + K_{23} - K_{12}\right)}{2}$$

$$P = 0.75\left[1 - e^{\left(\frac{K_{03}}{0.75}\right)}\right]$$

$$V_{03} = \frac{P(1-P)}{L\left(1-\frac{p}{0.75}\right)^2}$$

where L is the number of sites compared. Then the variance of K_{13}-K_{23} is given by V_{13}+V_{23}-$2V_{03}$. The variances, V_{13} and V_{23}, are calculated by taking the square of the standard error of K_{13} and K_{23}, respectively.

To determine if actin genes are evolving at different rates in different taxa, the mean number of non-synonymous substitutions (K_n), or amino acid substitutions, for all actin genes in each genome was calculated as described by Wolfe et al. (1987). In pooling the different genes to obtain the mean K_n for each taxa, the K_n value for each gene was weighed by its number of sites (L_n). The standard error of the mean K_n was calculated as the square-root of the mean variance

$$\overline{V}_K = \left(\sum_i L_i \right)^{-2} \sum_i L_i^2 V_{K_i}$$

where V_k and L_i are the variance of K and the L_n for the *I*th gene.

The Students t statistic was used to test the null hypothesis that the relative rate differences were not significantly different from zero (Zar, 1984). The t statistic was calculated as follows:

$$t = \frac{(\text{parameter estimate}) - (\text{parameter value hypothesized})}{\text{standard error of parameter estimate}}$$

where the parameter estimate is (K_{13}-K_{23}) and the parameter value hypothesized is zero.

When the null hypothesis is not rejected, there is the possibility of having committed a type II error, i.e., accepting the null hypothesis when it is false. Therefore, a power analysis was done to determine if the t-test had a high probability of detecting a statistically significant difference between the means (i.e., K_{13} and K_{23}). The power was estimated for the data using the STPLAN program (Brown et al., 1993) running on a Power MacIntosh computer. The options used were (1) normally distributed

outcomes, (2) 2-sample with unequal variances, and (3) 2-sided test. The input to the program was the number of substitutions (ie K_{13} and K_{23}), the sample size (i.e., the number of sites), the standard error of K_{13} and K_{23}, and the significance level (i.e., α=0.05).

In doing multiple t-tests, it is possible that the null hypothesis can be rejected with a certain frequency just by chance. Since the null hypothesis was rejected several times when determining the relative rate differences between taxa, the sequential Bonferroni test (Rice, 1989; Holm, 1979) was used to reduce the possibility of mistakingly rejecting the null hypothesis. The sequentional Bonferroni test reduces the α level depending on the number tests performed and thus reduces the probability of falsely rejecting the null hypothesis due to chance alone. The test is performed by first selecting a significance value (α), then replacing the test statistic by its corresponding P values and ranking these P values from smallest (P_1) to highest (P_k), where k is the number of, in this case, t-tests performed. If $P_1 \leq \alpha/k$, then the corresponding t-test is judged significant at the "table-wide" α level; if the inequalily is not met, then all t-tests are nonsignificant at the table-wide α level. If P1 $\leq \alpha/k$, then the second smallest P value (P_2) is analyzed for statistical significance. If $P_2 \leq \alpha/(k-1)$, then this test is also judged statistically significant at the α table-wide level and the third smallest P value (P_3) is analyzed for statistical significance. If P2 > $\alpha/(k-1)$, then the corresponding test and all other tests with larger P values are declared nonsignificant at the table wide α level. This is continued until P is no longer significant at the table wide α level.

Results

Part I: Description of actin clones

PCR amplification of actin clones

The amino and carboxyl ends of plant actins are extremely well conserved. Based on this fact, degenerate primers were designed using the plant actin sequences already available and used to PCR amplify actin genes from the angiosperm species, potato (*Solanum tuberosum*), tomato (*Lycopersicon esculemtum*), tobacco (*Nicotiana tabacum*), soybean (*Glycine max*), and maize (*Zea mays*). Figure 2 shows the phylogenetic relationships as well as some of the proposed times of divergence of these species. PCR was carried out at two different hybridization-temperatures (1 min @ 50°C or 2 min @ 54°C) for each species to ensure that as many different products were amplified. Fragments ranging in size from 0.8 to 2.5 kb were amplified for all five species. The number of fragments observed ranged from 4 to 10 for each species depending on the PCR conditions. In addition, given the same PCR conditions, the number of bands and their intensities also varied due to error associated with the PCR machine. Therefore, PCR was done at various conditions (see above) and on various days to maximize the number of PCR products available for cloning. These PCR products were then cloned into a phagemid vector and sequenced.

PCR was carried out using the enzyme *Taq* polymerase. This enzyme is known to have an error rate of about one in a thousand (Promega). Another enzyme, *Pfu* polymerase (Stratagene), which has a ten-fold lower error rate than *Taq* polymerase, was also used but failed to amplify any fragments using the same primers as above, regardless of the PCR conditions.

Table 4 shows the total number of unique actin clones isolated from each species. Some of the actin sequences in Table 4, such as Soy42, Soy117, Soy120 and

Table 4: List of actin sequences cloned from each species in this study. The size of introns and exons are indicated (in nucleotides). Number besides species name represents number of clones identified for that species. In some cases (indicated by > sign) the introns were not completely sequenced, so their size must be larger than indicated.

Species	Gene	Intron 1	Intron 2	Intron 3	Exon 2	Exon 3
Potato (8)	Pt42	73	absent	absent	394	602
	Pt46	85	174	453	394	614
	Pt65	79	absent	absent	394	617
	Pt66	107	220	305	394	614
	Pt79	109	220	243	394	614
	Pt82	>191	97	186	394	614
	PtII-2	absent	absent	absent	392	611
	Pt19	119	202	136	394	614
Tomato (6)	Tm32	63	absent	79	394	614
	Tm41	121	204	112	394	614
	Tm51	125	83	105	394	614
	Tm100	NA	NA	NA	NA	NA
	Tm52	NA	219	275	394	614
	Tm105	94	84	105	394	614
Tobacco (7)	Tb53	86	77	143	394	614
	Tb54	115	77	88	394	614
	Tb66	294	272	103	394	614
	Tb71	133	327	>105	392	614
	Tb93	110	131	>216	394	614
	Tb103	127	86	106	394	614
	Tb104	109	138	116	394	614
Soybean (12)	Soy42	NA	NA	NA	NA	NA
	Soy57	127	261	>280	394	614
	Soy58	>80	83	>200	394	614
	Soy69	94	301	81	394	614
	Soy70	>78	254	247	394	614
	Soy86	88	80	80	394	614
	Soy109	93	80	>129	394	614
	Soy115	114	229	>309	394	614
	Soy117	NA	>350	127	NA	614
	Soy118	139	319	>306	394	614
	Soy119	134	181	>350	394	614
	Soy120	143	336	NA	394	NA
Maize (11)	Mz31	NA	NA	NA	NA	NA
	Mz56	86	88	113	394	614
	Mz57	NA	NA	NA	NA	NA
	Mz63	>78	87	108	394	614
	Mz65	88	94	>81	393	606
	Mz72	110	NA	>160	NA	NA
	Mz81	>82	86	113	394	614
	Mz83	>74	84	>103	394	614
	Mz87	82	304	115	394	614
	Mz89	118	76	93	394	614
	Mz95	97	116	>374	394	614

Mz31 are truncated clones. While they have been established as representing unique clones based on their partial sequences, they were not analyzed further. Presumably, full length copies of these genes are present in the respective genomes. Clones Tm100 and Mz57 are also unique and appear to be full length actins, but were not fully sequenced. The maize actin gene, Mz72, appears to be a pseudogene since it has a large deletion in the 3'end of exon 2 and in the 5'end of exon 3. This made it difficult to determine the location of the intron borders. In some instances, the sequence of the first and third intron were not completely determined.

A number of these actin genes are identical to actin genes previously cloned and sequenced (Table 5). For example the potato clones PtII-2 and Pt19 are identical in sequence to the processed pseudogene, PoAc99 (Drouin and Dover, 1987), and PoAc97 (Drouin and Dover, 1990) respectively. In addition, Soy86 is nearly identical to a previously cloned actin gene from soybean, SAc3 (Shah et al., 1982). It is surprising that, out of the 42 actin genes cloned and sequenced in this study, only three were identical to previously cloned actins. This is especially so for soybean and maize where 12 and 11 different actins were identified respectively, and yet there was only one match to a previously cloned actin. Similarly, out of the 8 different actins isolated for potato, only two matched the five actins previously identified. These results indicate that the actin multigene family in these plants may be relatively large.

Actin gene copy number

It is possible to roughly estimate the number of actin genes present in the soybean and potato genomes using the mark-recapture method ecologists use to estimate population size. The formula is given by

Taxa	Unique PCR Clones	Genes in databanks	Matches to previous clones	Total
Potato	8	6	2	12
Tomato	6	0	0	6
Tobacco	7	1	0	8
Soybean	12	2	1	13
Maize	11	1	0	12
total	**44**	**10**	**3**	**51**

Table 5: Numbers of actin genes cloned in this study from five plant species. The number of previously cloned actin genes for these species is shown as well as the number matches to actins cloned in this study.

$$N = \frac{nM}{x}$$

where N is the total number of actin genes in the genome of the species under study, n is the number of previously identified actin genes for the same species, M is the total number of actin genes cloned in this study, and x is the number of matches between the actin genes previously cloned and the actin genes cloned from this study. The standard error of N is given by

$$N\sqrt{\left[\frac{(N-M)(N-n)}{nM(N-1)}\right]}$$

(Ricklefs, 1990). Accordingly, the number of actin genes in potato and soybean is estimated to be 24±12 and 24±17, respectively. Since no matches were found for other species, the actin copy number can not be estimated using this method. These numbers are in agreement with those obtained using southern analysis (Thangavelu et al., 1993).

Location and number of introns

The location and the number of introns in angiosperms actin genes is known to be conserved. All angiosperm genes cloned and sequenced to date contain three introns located in the exact same locations along the amino acid sequence. Table 4 lists the actin sequences cloned in this study. Because some of the clones have not been completely sequenced or in some cases they are truncated actin sequences, information regarding size and number of introns is not available. With a few exceptions, all the angiosperm actin clones contain three introns. The first intron is located between amino acids 20 and 21, the second intron is located after the first base of the codon for amino acid 132,

and the third intron is located between amino acids 356 and 357. This number system is based on the first exon of angiosperm actin sequences having 20 amino acids.

Intron loss in angiosperm actin genes

While no new intron positions have been found among these actin clones, some of these genes appear to have lost one, two or all three introns. The potato actin clone PtII-2 contains no introns and is identical in sequence to the potato processed pseudogene (Drouin and Dover, 1987). Processed pseudogenes originate when a mRNA is reverse transcribed into cDNA (lacking introns) which then integrates back into the genome (Li and Graur, 1991). The tomato actin clone, Tm32, has lost the second intron while the potato actin clones, Pt42 and Pt 65, have lost their second and third introns. In all cases the removal of the introns is precise, leaving the coding sequence in frame. Based on the sequence similarity of the introns as well as the coding sequence, intron loss in Pt42 and Pt65 most likely represents a single event in the ancestor of these two sequences. Whether or not intron loss in these tomato and potato genes represent independent events will be discussed below in the context of phylogenetic analysis.

Size of introns

In angiosperms, there is a limit on how small intron sequences can be and still retain their ability to be efficiently spliced out of the pre-mRNA (Goodall and Filipowicz, 1990). Introns smaller than 68 nucleotides are inefficiently spliced out, or not at all. The failure of short introns to be spliced in plants and vertebrates could be the result of steric hindrance preventing the assembly of the full complement of small nuclear ribonucleoprotein particles (snRNPs) and other factors on the intron. In the plant actin genes cloned in this study, all introns, with the exception of the first intron

of the tomato actin clone Tm32, are larger than 68 nucleotides. The average intron lengths are >109nt, >172nt and >169nt for the first, second and third introns respectively. These numbers are underestimates because in some cases the complete intron sequence was not determined. The size of the introns ranged from 63 nt to 453 nt. It is possible that some of the introns which were not completely sequenced are longer than 450 nucleotides. As already mentioned, Tm32 contains a stop codon in the second exon and has lost the second intron. Thus, the fact that it also contains an intron which is too short to be spliced out also points to it being a pseudogene. However, short introns have been found in both plants and animals (Green, 1986). These could either be inefficiently spliced or could contain special structural features that bypass the requirement for one or more splicing factors, or have a structure that allows optimal packing of the splicing factors on the intron.

Conservation of intron splice regions

Angiosperm introns contain consensus sequences at their 5' and 3' ends which are needed for splicing out of the nuclear mRNAs (White et al., 1992; Perlman et al, 1990). With some exceptions, all the introns of the actin clones contained the typical invariant dinucleotides at their 5' and 3' ends. The 5' ends of these introns begin with GT and terminate at their 3' ends with the AG dinucleotide. However, the second intron of Soy115 and Mz95 as well as the third intron of Pt46, Tm51, Tb103 start with the dinucleotide GC. At the 3'end, clones Tb104, and Soy69 end with the dinucleotide AT. Whether or not these changes affect the splicing out of these introns is not known.

No conservation of intron sequences

Evolutionary rates of nucleotide substitution in noncoding regions are typically significantly higher than those of coding regions (Li and Graur, 1991). This is

reflective of the relatively low selective constraints operating on noncoding relative to coding DNA. It is therefore not surprising that there is not much conservation in both the size and nucleotide sequence between the respective introns of the different actin sequences. However, there are some exceptions. In these cases the genes have either diverged from one another relatively recently and not enough nucleotide changes have accumulated or gene conversion has homogenized the sequences (see below).

Size of coding region conserved

The size of the actin coding region is known to be well conserved among eukaryotes (reviewed in Shetterline and Sparrow, 1994). Table 4 lists the size of exons 2 and 3 for all the clones identified in this study. The size of these two exons is very conserved, exon 2 and exon 3 being 394 and 614 nucleotides long, respectively. However, some clones vary a little in size. The potato clone Pt42 has a 12 base pair deletion in exon 3 which does not affect the reading frame. In addition, the potato clone Pt65 has a three nucleotide insertion in exon 3 in the same region as the deletion in Pt42 and also does not affect the reading frame. This deletion and insertion is in a variable region of the coding sequence (see Figure 4) so it is not known if the products of these two genes are active. As mentioned above, these two genes also lack introns 2 and 3.

The tobacco clone Tb71 has two single nucleotide deletions in exon 2 which affect the reading frame by producing two stop codons. The maize clone Mz65 has a single base deletion in exon 2 and various others in exon 3 which affect the reading frame. Mz65 also has the highest G+C content of all the plant actin sequences known to date (see below). Thus both these actin clones are likely pseudogenes. The clone Mz65 will not be included in the phylogenetic studies because too many gaps needed to be introduced into the alignment.

Amino acid alignment

The amino acid sequences for most of the actin clones identified in this study were deduced from their respective nucleotide sequences and aligned with other previously identified plant actin amino acid sequences as well as an actin sequence from *Volvox carteri* (Figure 4). As mentioned already, the size of these proteins is very well conserved and very few gaps are needed to generate an unambiguous alignment. The amino acid alignment also gives a graphic representation of the high degree of conservation of plant actins. The plant actin amino acid sequence has been demarcated into blocks A through H. These areas have been shown to be highly conserved in animal actin sequences and represent regions in the animal actin proteins responsible for actin-actin and actin-binding-protein interactions (Reece et al., 1992). From the alignment, it is obvious that most amino acid changes occur for the most part outside these blocks, especially in the areas between blocks F and G. The higher degree of conservation in these blocks suggests that the same areas of the plant actin sequences are also responsible for protein-protein interactions.

Variation in amino acid and nucleotide sequences

Table 6 shows the pairwise comparisons of percent amino acid and nucleotide differences between 53 plant actin genes (including the *Volvox* actin gene). Actin nucleotide sequences can vary as much as 21-25% both within and between species. This suggests that the actin multigene family may be equally old in all these lineages. The fact this level of variation is also observed between the *Volvox* actin gene and all the other plant actin genes suggests that variation in nucleotide sequence between some plant actin genes has reached a saturation point. This observation is consistent with the

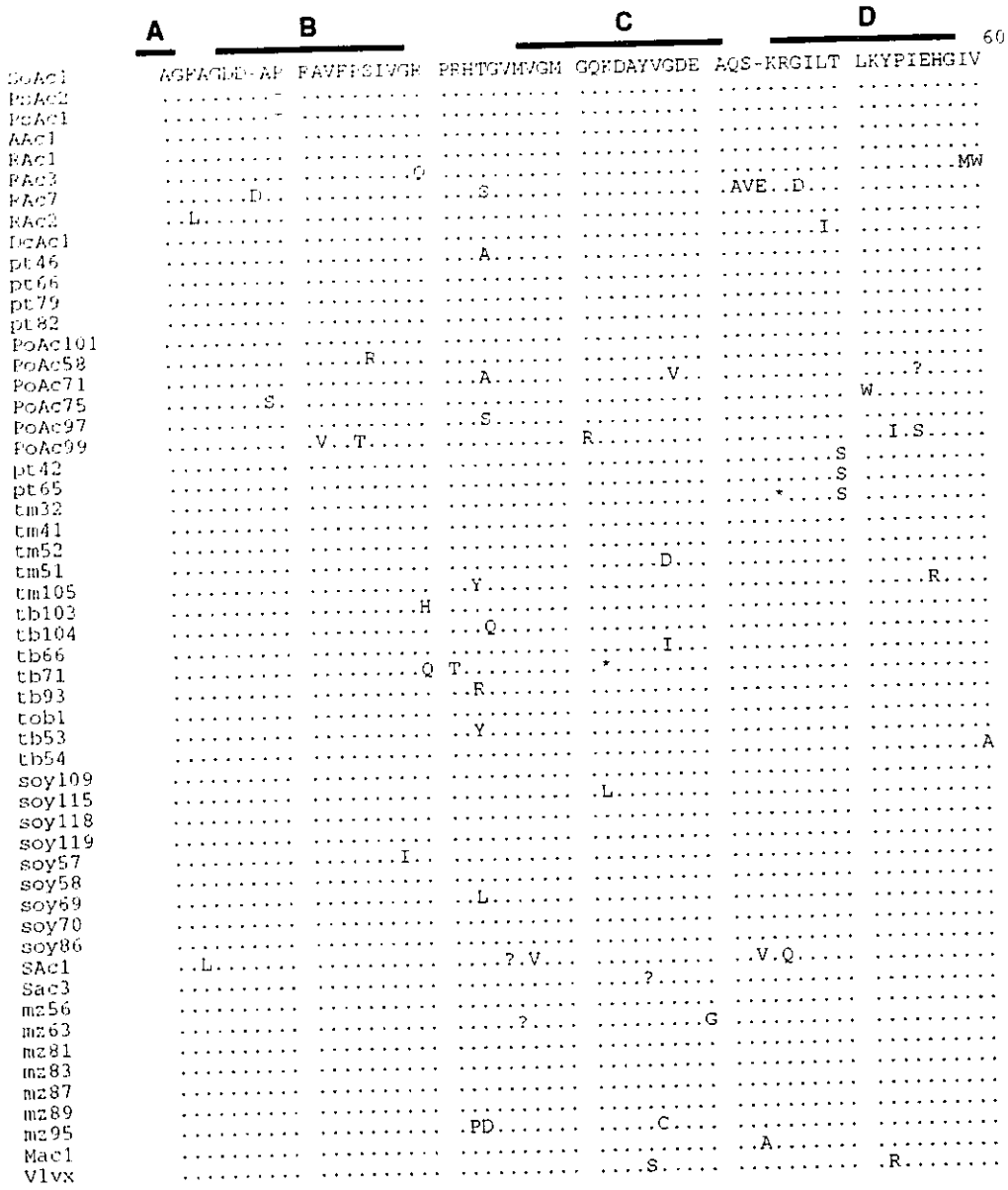

Figure 4: Amino acid alignment of 54 plant actins. Regions under the solid line labelled A through H represent regions of high sequence conservation with animal actin genes. Dots represent identical amino acid residues to top sequence. Stop codons are represented by the * symbol. Unknown amino acids are denoted by a question mark (?), and introduced gaps are denoted by hyphens (-). In cases where a frame shift mutation occurred, the frame was corrected by inserting N in the nucleotide sequence and these codons are denoted by X in the amino acid alignment.

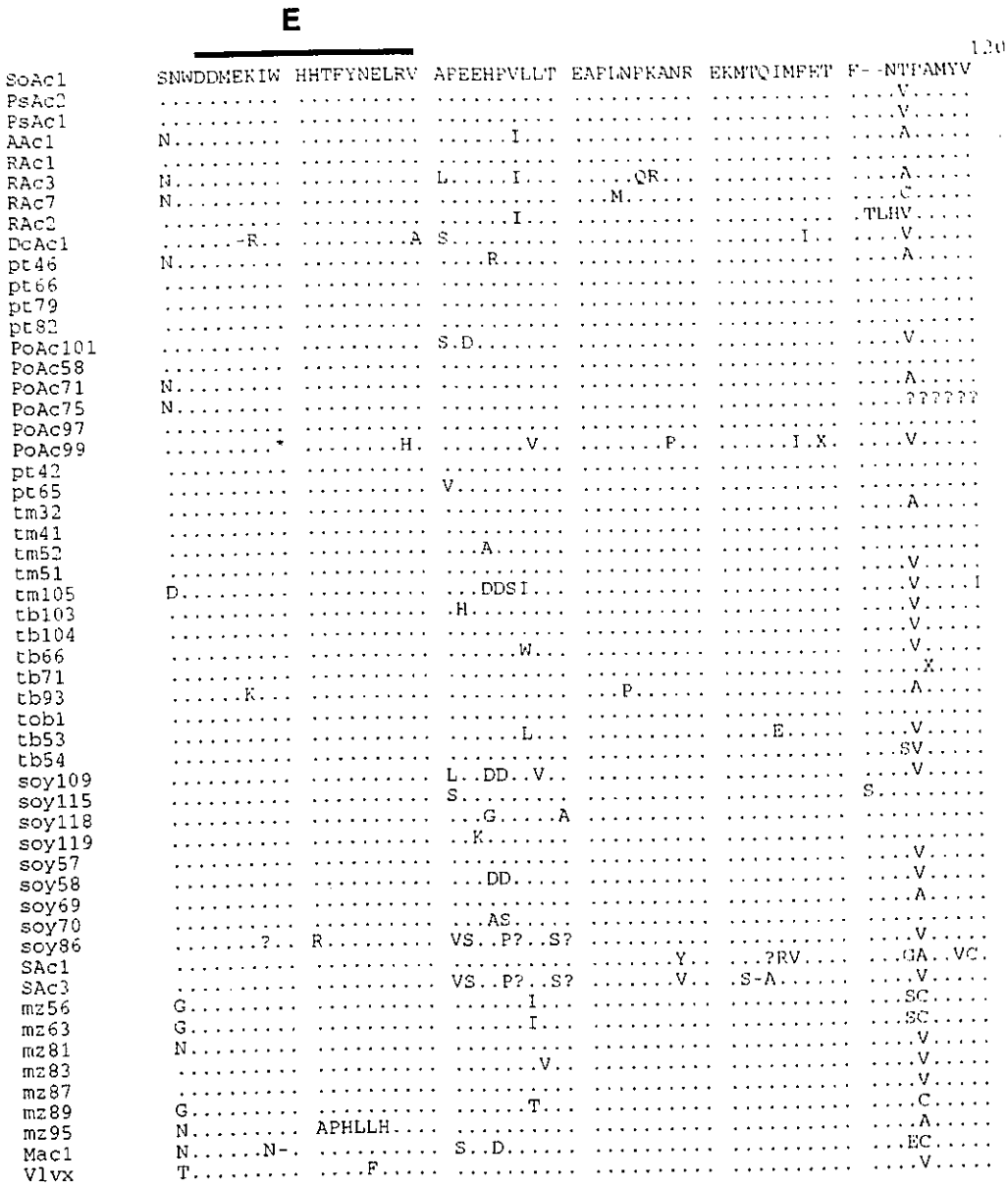

Figure 4: Con't

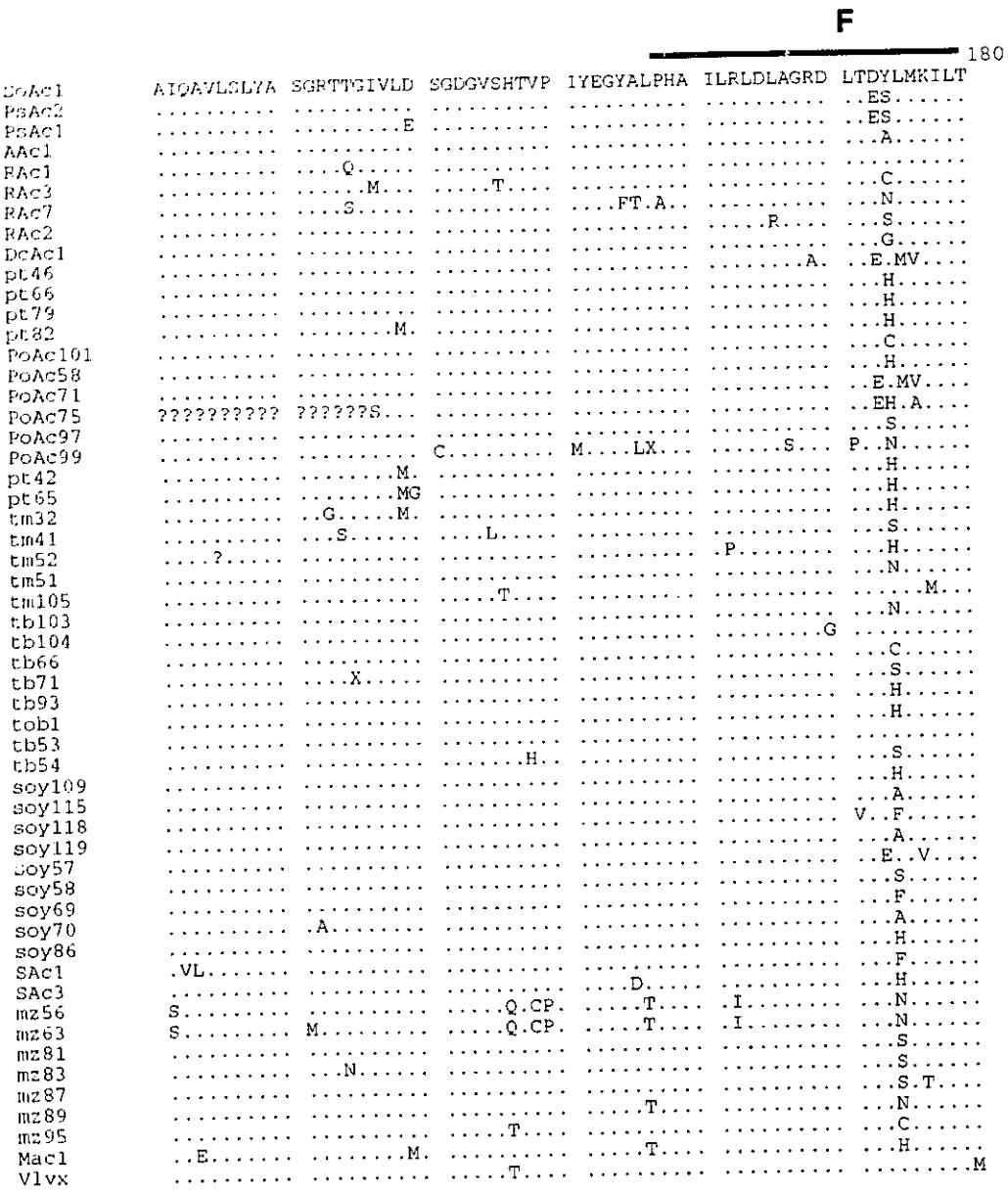

Figure 4: Con't

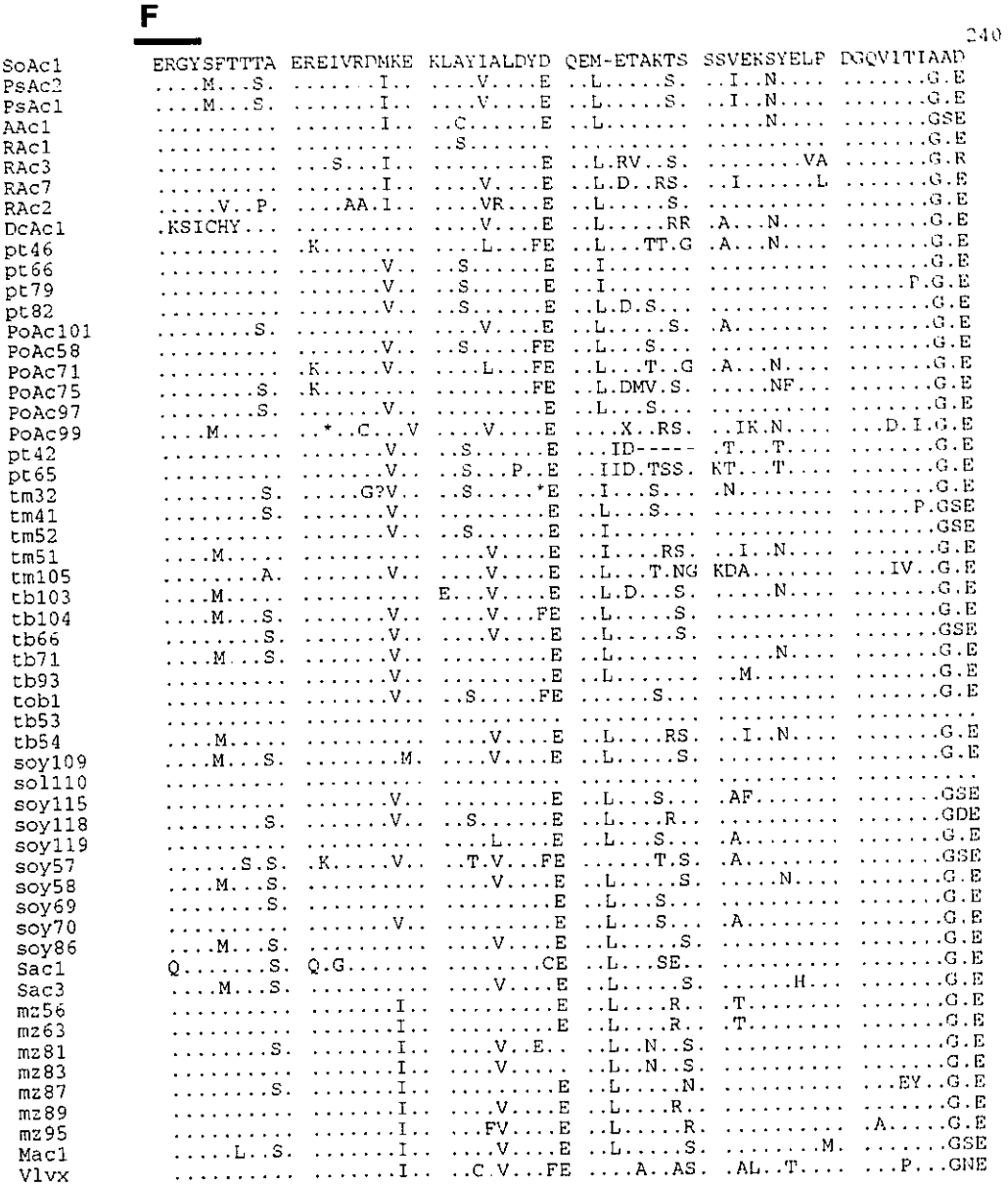

Figure 4: Con't

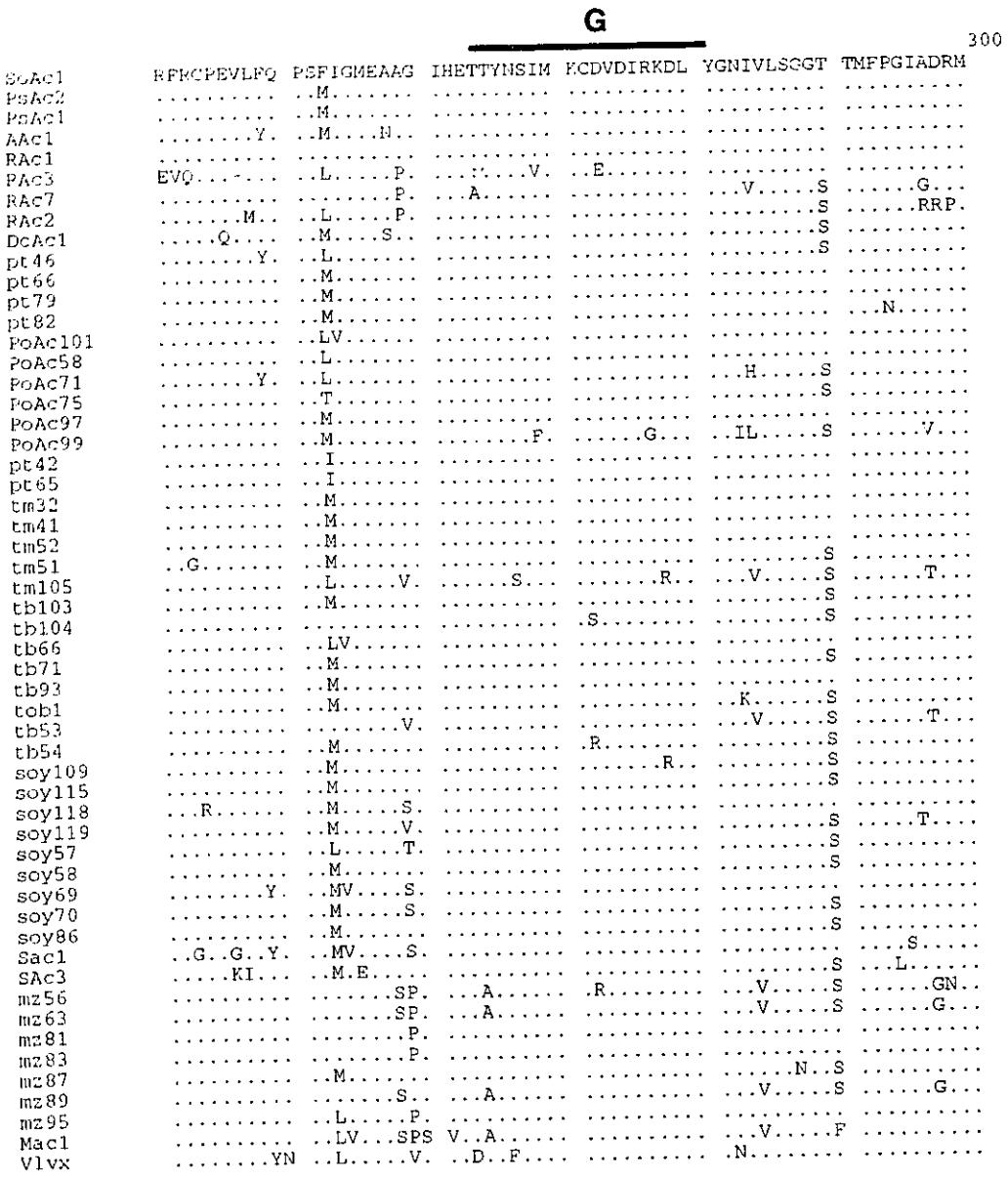

Figure 4: Con't

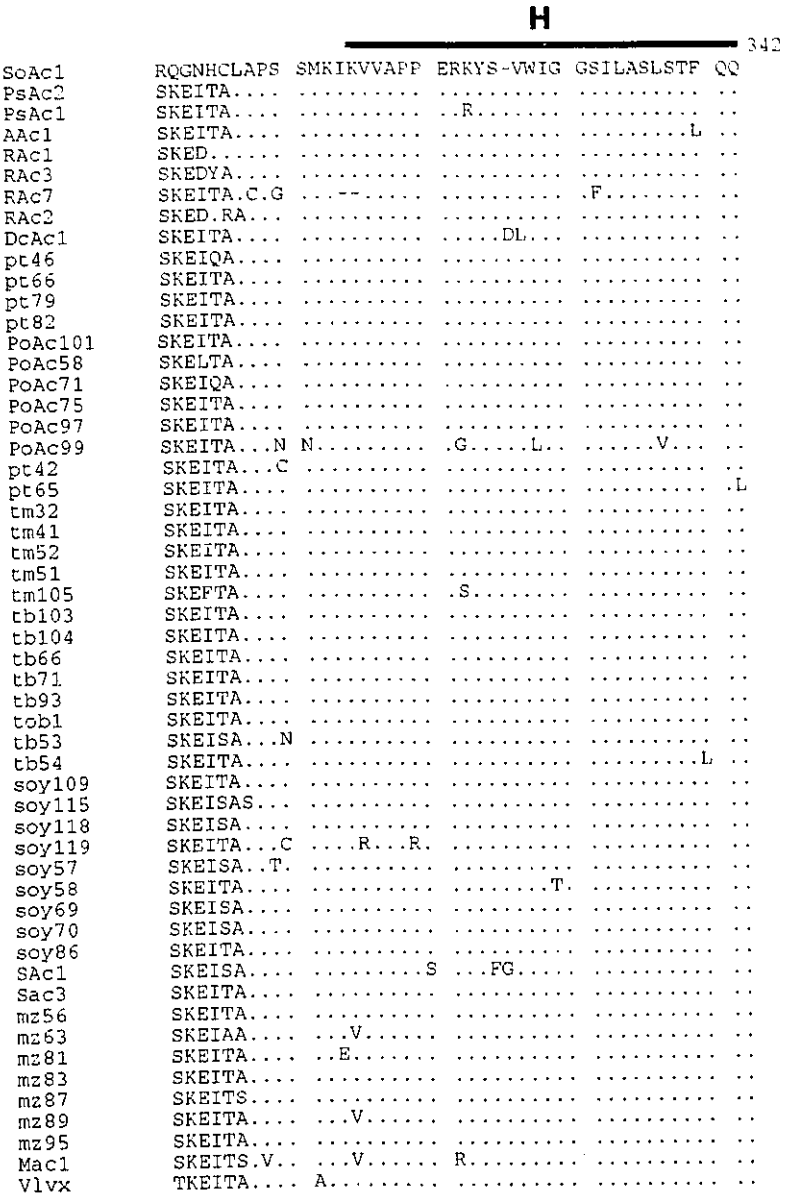

Figure 4: Con't

Table 6: Pairwise comparisons of percent amino acid differences (above diagonal) and percent nucleotide differences (below diagonal) between 53 plant actin genes. Numbers in the top row refer to the corresponding numbered genes in the first column. Boxed areas represent pairwise comparisons within a single species.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53
1. Vivx		10	8	9	8	9	13	14	13	12	8	9	10	9	16	9	10	10	8	9	9	14	10	10	10	8	9	13	13	9	9	9	9	10	9	10	10	10	9	9	11	10	9	9	10	12	9	10	11	10	10	15	12
2. SoAc1	23		7	7	7	2	10	11	11	10	6	5	9	5	15	6	8	8	5	5	5	12	10	7	7	5	6	10	10	5	7	6	7	6	6	8	7	6	6	7	8	7	7	8	7	12	6	7	9	7	6	12	9
3. PsAc2	25	21		1	5	5	8	10	8	7	4	4	6	3	11	5	7	6	4	4	4	10	7	5	4	3	3	9	9	4	3	3	4	3	4	4	6	4	5	3	5	6	6	7	3	10	5	4	8	6	6	11	6
4. PsAc1	25	21	0		5	6	9	11	9	7	4	5	7	3	12	6	7	7	4	5	5	10	8	6	4	4	4	10	10	5	4	4	4	3	4	4	6	4	5	3	5	6	6	8	6	11	5	7	9	5	6	11	9
5. AAc1	23	21	18	19		6	10	9	10	8	6	5	6	4	14	6	7	6	4	4	4	10	8	6	6	5	5	9	8	5	5	6	5	5	4	6	6	5	4	6	7	6	6	7	6	11	5	6	8	5	5	11	8
6. HAc1	22	9	19	20	19		8	10	10	9	5	3	7	4	14	4	6	7	3	3	3	11	8	6	5	4	5	9	9	3	5	5	5	5	4	6	5	5	4	5	6	5	5	7	6	11	5	6	8	5	5	11	8
7. HAc2	25	20	19	20	21	19		12	12	12	8	8	11	8	17	10	12	11	9	9	9	12	11	9	8	7	8	11	11	9	9	8	8	9	9	9	10	9	10	9	9	10	9	10	8	12	9	9	10	9	9	14	11
8. RAc3	27	23	20	20	21	21	21		13	13	10	9	12	9	19	11	12	12	9	10	10	14	10	11	11	9	9	12	12	10	11	11	10	10	9	12	11	10	10	11	11	11	11	13	11	14	10	11	13	11	10	15	13
9. RAc7	29	23	23	23	23	23	24	23		12	9	9	10	8	16	10	12	11	9	9	9	10	11	5	9	8	8	9	9	8	9	9	8	9	9	8	9	11	10	10	11	11	10	10	9	12	11	10	10	11	11	9	15	11
10. DcAc128		25	18	18	22	23	24	23	23		6	8	9	7	15	9	10	9	7	8	8	12	11	8	8	8	8	12	11	7	6	7	8	7	8	7	9	8	8	6	7	9	7	9	6	13	8	7	10	7	8	14	9
11. PoAc101	27	22	17	18	21	20	21	21	24	19		4	7	3	13	5	6	6	4	4	4	8	7	6	5	4	4	9	9	5	4	4	2	5	4	5	5	4	5	4	5	5	4	5	4	6	4	9	4	5	8	4	4	10	7
12. PoAc58	26	21	19	19	18	19	21	23	25	23	21		6	2	14	3	4	7	1	2	2	10	7	5	5	4	5	9	8	2	5	4	4	4	3	6	4	3	2	5	5	4	5	6	5	9	4	5	8	4	4	10	7
13. PoAc71	27	24	21	21	20	22	23	23	27	23	21	21		6	15	7	9	2	6	7	7	12	9	7	7	7	7	10	10	6	7	6	7	6	6	8	8	7	7	7	8	8	7	6	7	11	7	8	9	7	6	13	10
14. PoAc97	26	19	18	18	17	17	20	22	24	21	19	10	20		13	4	5	6	2	2	2	9	7	5	3	3	3	8	8	2	4	3	3	2	2	5	3	2	3	4	4	3	4	6	3	9	3	4	7	4	3	9	6
15. PoAc99	26	23	17	18	21	21	21	23	24	21	21	21	23	20		14	16	15	13	13	13	19	17	13	13	13	14	17	16	12	11	13	14	12	13	12	14	14	14	10	12	15	13	15	12	17	13	13	15	14	14	19	14
16. pt42	25	22	20	20	19	20	23	22	25	22	20	17	19	18	22		2	8	2	3	2	11	9	7	6	5	6	9	9	3	6	6	5	5	4	7	3	5	3	6	7	5	5	7	6	11	6	6	8	5	5	12	8
17. pt65	26	22	20	20	19	21	23	22	25	22	21	16	19	18	22	2		9	3	4	4	12	10	8	7	7	7	11	10	4	7	7	7	6	5	8	4	6	4	7	8	8	6	6	6	11	6	7	9	7	6	12	9
18. pt46	27	23	21	22	20	22	23	23	27	23	21	21	3	20	23	20	20		7	7	7	12	9	8	7	7	7	11	11	7	6	7	7	7	7	7	8	8	8	7	8	8	6	6	6	11	6	7	9	7	6	12	9
19. pt66	25	18	17	17	17	15	20	21	23	20	19	16	18	15	20	16	17	19		0	1	10	7	5	4	4	4	8	8	1	4	4	4	3	2	5	2	3	1	4	5	3	4	6	4	10	4	3	7	4	3	10	6
20. pt79	24	18	17	17	17	16	20	21	23	21	19	16	19	16	20	17	17	19	1		1	10	7	5	4	4	5	9	8	1	4	4	4	3	2	5	3	3	2	5	5	3	4	7	5	10	4	5	8	4	4	10	7
21. pt82	27	23	20	20	19	20	23	22	26	23	19	16	18	16	21	13	13	19	16	17		9	7	5	4	4	5	9	8	2	4	4	4	3	2	5	3	3	2	5	5	3	4	7	5	10	4	5	8	4	4	10	6
22. MAc1	27	21	21	21	21	20	23	22	19	23	21	22	24	22	23	24	24	23	21	21	24		12	8	10	9	9	11	10	10	11	10	8	10	10	11	10	10	10	11	10	10	11	11	10	14	10	10	12	10	11	15	12
23. mz95	24	21	20	20	22	21	22	17	23	24	22	23	26	23	23	24	25	25	21	21	24	22		8	8	6	6	11	11	8	8	8	7	8	7	9	9	8	8	8	9	9	8	10	8	12	7	8	10	9	8	14	11
24. mz89	24	19	19	19	19	19	21	21	17	21	21	21	22	20	21	21	22	22	20	20	21	15	20		6	5	5	4	4	5	5	5	5	6	5	6	7	6	6	5	6	7	6	8	6	9	6	6	7	6	6	12	8
25. mz87	24	20	18	18	18	19	20	21	23	22	18	21	23	20	20	23	24	23	19	20	21	20	21	19		4	4	9	9	4	5	4	5	4	4	5	5	4	5	5	5	5	5	7	4	10	5	5	7	6	5	11	7
26. mz83	23	18	17	17	19	16	17	20	22	21	19	21	23	19	20	21	21	23	18	18	22	18	19	17	17	7		8	8	5	5	4	4	5	5	6	6	4	5	5	5	5	6	7	4	10	5	5	7	6	6	11	7
27. Mz81	23	19	17	17	18	18	17	20	21	21	19	20	22	19	20	21	21	21	19	19	22	18	19	17	17	7		8	8	5	5	4	4	5	5	6	4	5	5	5	5	6	7	4	10	5	5	7	6	6	11	7
28. Mz63	26	20	23	23	22	21	23	24	21	24	25	25	27	24	25	25	25	26	24	24	25	18	21	8	21	19	19		2	8	9	9	9	9	9	8	10	9	9	9	10	10	9	11	9	12	9	10	10	9	8	15	12
29. Mz56	25	20	22	22	22	20	22	22	21	25	24	25	27	24	25	25	25	26	23	23	25	18	21	10	20	17	16	3		8	9	8	9	9	8	8	10	9	9	9	10	10	9	11	9	12	9	10	10	9	8	15	12
30. Tob1	27	22	21	21	20	20	22	24	26	23	20	16	19	17	22	11	12	20	17	17	11	24	23	22	22	21	21	25	25		4	4	4	3	3	5	3	3	2	4	5	4	4	6	4	10	4	4	7	4	3	10	6
31. tb103	24	20	15	16	19	17	18	19	23	19	19	18	21	17	9	19	19	20	16	16	19	20	20	18	19	17	17	22	21	20		4	5	3	4	3	6	5	5	2	4	6	5	7	3	10	5	3	7	5	5	11	6
32. tb104	25	21	16	16	20	21	22	22	22	21	19	20	20	19	20	20	20	21	18	19	19	23	22	21	19	21	20	24	24	19	18		4	3	4	4	5	4	5	3	9	5	4	6	5	7	5	5	10	7
33. tb66	28	23	17	17	22	22	21	21	25	19	7	21	22	20	22	21	21	22	19	19	19	22	21	22	20	21	20	25	24	21	20	19		4	4	6	5	3	4	5	5	5	6	5	4	9	4	5	7	5	5	10	6
34. tb71	25	19	17	17	16	17	20	22	24	21	20	11	21	5	19	19	19	21	16	17	18	22	23	20	20	19	19	23	23	17	17	19	20		4	4	5	3	3	4	5	4	5	7	5	9	3	5	7	4	4	10	7
35. tb93	25	19	18	18	18	16	20	22	23	20	20	18	19	17	20	17	18	19	6	7	17	21	21	20	20	18	19	23	24	17	17	20	20	18		5	4	4	4	5	3	4	4	4	5	4	8	4	4	10	7
36. tb54	25	21	17	17	20	19	19	20	23	21	19	18	21	17	13	20	20	21	18	19	20	21	21	20	21	19	19	23	23	20	10	18	20	17	18		7	6	6	3	5	7	6	8	3	11	6	5	8	6	6	12	7
37. tm32	27	24	21	21	21	22	24	24	27	24	22	19	21	20	24	10	11	21	19	19	16	25	25	23	24	23	23	26	26	14	21	22	23	21	18	22		4	3	6	6	5	6	7	6	11	5	6	9	6	5	11	8
38. tm41	26	19	19	19	17	17	21	23	25	22	20	11	21	2	21	18	18	21	16	16	17	22	23	20	21	19	19	25	24	17	18	21	20	6	17	18	20		3	5	5	4	5	6	4	10	4	5	8	4	4	10	7
39. tm52	25	18	17	17	17	15	20	21	23	20	19	17	18	16	21	17	17	19	2	2	17	20	21	20	20	18	18	19	24	23	17	17	18	19	17	7	19	19	17		5	5	3	5	6	5	10	5	5	8	4	4	11	7
40. tm51	24	20	14	14	18	18	18	19	22	19	18	18	21	16	7	19	19	21	17	18	19	20	20	18	18	17	17	22	22	19	6	17	19	16	18	10	21	17	18		4	6	5	7	3	10	5	4	8	5	5	11	6
41. soy109	26	20	9	9	19	19	21	20	24	19	18	19	21	18	18	20	20	21	17	18	20	21	22	19	19	18	18	22	22	20	16	16	18	17	19	17	22	19	18	16		6	6	7	2	9	5	3	8	6	5	11	5
42. soy118	25	19	19	19	18	18	20	22	26	23	21	17	21	15	21	19	19	21	15	16	19	23	22	21	20	19	19	24	24	20	18	23	21	16	17	20	20	17	16	18	20		6	6	5	11	3	6	8	5	4	9	8
43. soy119	25	19	18	18	18	18	21	22	24	21	20	17	19	16	21	18	18	19	15	15	17	21	23	20	18	18	18	23	24	18	19	20	21	17	17	20	20	17	16	18	18	17		7	5	10	5	5	7	4	4	11	7
44. soy57	26	22	19	19	23	21	22	23	26	20	17	22	20	21	21	22	22	20	20	20	21	22	23	21	21	21	20	24	24	22	20	20	18	21	20	20	22	22	21	19	18	22	20		7	11	7	7	8	6	6	13	9
45. soy58	25	21	8	8	19	19	21	20	23	18	18	19	22	18	18	20	21	22	18	19	20	21	21	19	19	17	17	23	22	21	16	16	18	17	19	17	21	19	19	16	6	20	19	18		9	5	3	7	6	5	11	5
46. tm105	27	24	19	19	20	22	23	22	24	22	21	23	21	21	22	21	22	21	21	21	20	23	24	22	21	22	21	25	25	21	19	13	20	21	21	20	23	23	21	20	18	23	21	22	18		10	10	7	10	9	16	12
47. soy69	26	22	20	20	18	20	21	21	26	23	22	18	19	17	22	19	19	19	17	17	18	24	21	22	20	20	20	25	24	20	19	21	21	17	17	20	19	17	17	19	20	10	17	21	21	23		5	7	5	4	7	7
48. soy86	27	21	9	9	19	20	21	22	24	19	18	19	21	18	18	20	20	21	18	19	21	22	23	19	20	18	18	23	22	21	16	16	18	17	19	17	22	19	19	16	3	20	19	19	8	19	20		8	6	5	11	3
49. tb53	26	23	17	18	20	22	22	23	21	20	21	21	20	21	21	22	21	20	20	20	20	23	23	21	20	21	20	24	24	22	18	12	20	20	21	20	23	22	20	19	18	22	21	21	18	8	22	18		8	7	13	10
50. soy115	26	21	19	19	19	20	22	22	25	22	21	19	20	17	21	19	19	20	16	16	18	21	22	20	20	19	19	23	23	19	19	21	20	17	18	21	20	17	16	18	18	17	10	21	19	22	17	20	21		3	11	8
51. soy70	27	21	19	19	18	20	22	22	25	23	21	18	19	17	22	18	18	19	16	16	18	21	22	21	20	19	19	23	23	19	19	20	21	17	17	21	20	17	16	19	19	16	10	21	19	22	17	19	21	4		10	7
52. SAc1	28	24	22	22	22	22	22	23	24	28	24	24	21	21	20	24	22	22	21	19	19	21	26	24	24	22	22	23	27	26	22	22	23	23	20	19	21	21	20	20	21	22	14	20	24	24	25	6	23	24	20	20	
53. SAc3	28	22	10	10	21	21	22	23	26	20	20	20	22	20	19	22	21	22	19	20	21	23	24	21	21	20	19	24	24	22	17	17	19	19	21	18	24	20	20	17	5	21	20	20	9	20	21	2	19	21	21	24	

actin protein being under strong functional constraints and thus limiting the amount of possible nucleotide change.

At the amino acid level, actin sequences in plants (not including *Volvox*, Sac1 and PoAc99) vary from 0-14%. Within a species, the rice actin genes are unique in that none of them differ from one another by less than 8%. This could be the result of having too few genes in rice available for comparison (i.e. limited sampling of rice genes). Alternatively, if the actin multigene family in rice is small, it could mean that most of the actin genes are ancient. The actin multigene family in rice is thought to be composed of at least 8 members. The variation in amino acid sequences is 0-9% in potato (not including PoAc99), 2-12% in maize, 3-8% in tobacco, 3-11% in tomato, and 2-9% in soybean (not including Sac1). The low variation observed in tobacco and tomato could be the result of having fewer genes for comparison.

In every species there are actin genes which are more similar, in both amino acid and nucleotide sequence, to actins of other species than they are to those within the same species (also see phylogenetic trees below). For example, the tomato actin genes vary from each other by more than 17% in their nucleotide sequence, yet Tm52 and Tm41 differ by only 2% from Pt79 and PoAc97, respectively. Similarly, some of the rice actin genes differ by as little as 3% in their amino acid sequence to actins of other species. The fact that there are genes within each species which are more similar to actins from other species suggest that many of these actin genes may be orthologous (see below).

At the amino acid level, the highest sequence divergence (up to as much as 19%) occurs with comparisons to the potato processed pseudogene, PoAc99. This is expected since this gene is no longer under any functional constraints. The difference between plant actin amino acid sequences and the *Volvox* actin varies from 8-16%. A

high amino acid sequence divergence (9-16%) is also observed with comparisons to the soybean actin gene, Sac1. This suggests that Sac1 may also be a pseudogene.

Part II: C+G content and codon usage

G+C content of genes

The G+C content of angiosperm actin genes varies within a relatively narrow range between 42% (Pt82) and 65% (Mz65) (Figure 5) and dicots have lower G+C content than that of monocots. Out of 14 monocot genes looked at, 13 of these had higher G+C contents than all of the dicot actin sequences. The average G+C content for dicot actins is 46% ± 2% whereas in monocots the average is significantly higher (54% ± 4%). Within dicots there is not any statistical difference in the G+C content of actin genes between the different species. The same can be said for monocots, although the sample size is smaller. Interestingly, the gene with the highest G+C content is Mz65, which appears to be a pseudogene due to its many frameshift mutations. This suggests that in the absence of selection, its G+C content has drifted to the average G+C content of the maize genome (Campbell and Gowri, 1990).

Effect of G+C content on codon usage

Since there is a significant difference in the G+C content between dicot and monocot actins, we wished to determine if this difference has some effect on the codon usage patterns of dicot and monocot actins. If such a difference exists it will best show up in the third position of codons since it is degenerate and therefore is less constrained by selective forces. Figure 6 shows the frequency of G+C at the third codon position plotted against the frequency of A at the same position. The G+C content at the 3rd position of codons in plant actin genes is distributed over a large range. The frequency of G+C at the third position of codons can be as low as 32% for dicot actin genes and

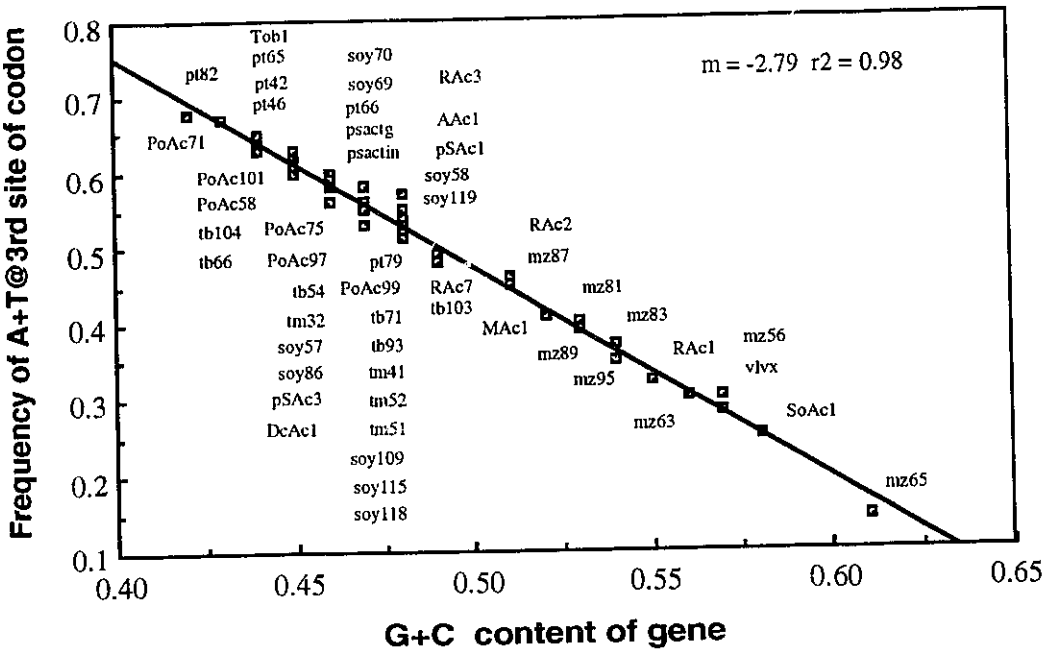

Figure 5: Variation in the frequency of G+C content among plant actin genes.

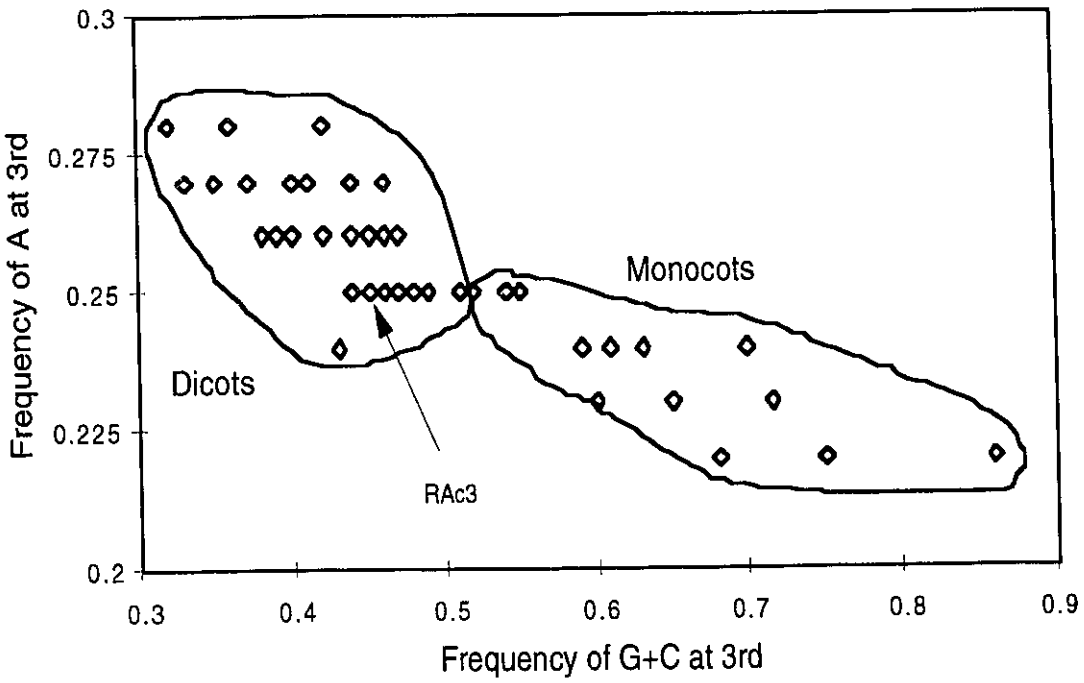

Figure 6: Codon usage patterns among plant actin genes as reflected by their G+C content at the third codon position.

as high as 86% for monocots. In general, dicot actin genes have a bias for codons ending in A or T while monocot actins have a preference for codons ending with G or C. The average G+C content at the third position for dicots and monocots is 43% ± 5% and 63% ± 10%, respectively. This codon usage bias determined from the G+C content of the third position of codons is expected to affect only synonymous substitution patterns and not the non-synonymous substitution patterns.

Effect of G+C content on amino acid usage

To determine if the differences in codon usage between dicot and monocot actin genes caused by differences in G+C content affect the non-synonymous substitution pattern, and thus the amino acid sequence, the effect of the G+C content of the gene on the G+C content of the individual codon positions was studied. Figure 7 shows that there is no correlation between the G+C content of the gene with the G+C content at the first and second position of codons. Thus, the G+C content of the genes, even between monocots and dicots, has no effect on the non-synonymous substitution patterns. As expected, the G+C content at the third position of codons is directly correlated with the overall G+C content of the coding region and thus the synonymous substitution patterns are not independent of G+C content.

These results indicate that using the synonymous rates of substitution to construct distance-based phylogenetic trees or including the third position of codons to construct parsimony trees will result in biases. It is likely that genes may cluster on the tree based on shared G+C content at the third position of codons, or because they have smaller synonymous distances associated with them as a result. Therefore phylogenetic trees will be constructed either by using only the first and second position of codons or non-synonymous substitutions. This will ensure that genes will cluster on the tree because they are of common descent rather than by spuriously shared G+C contents.

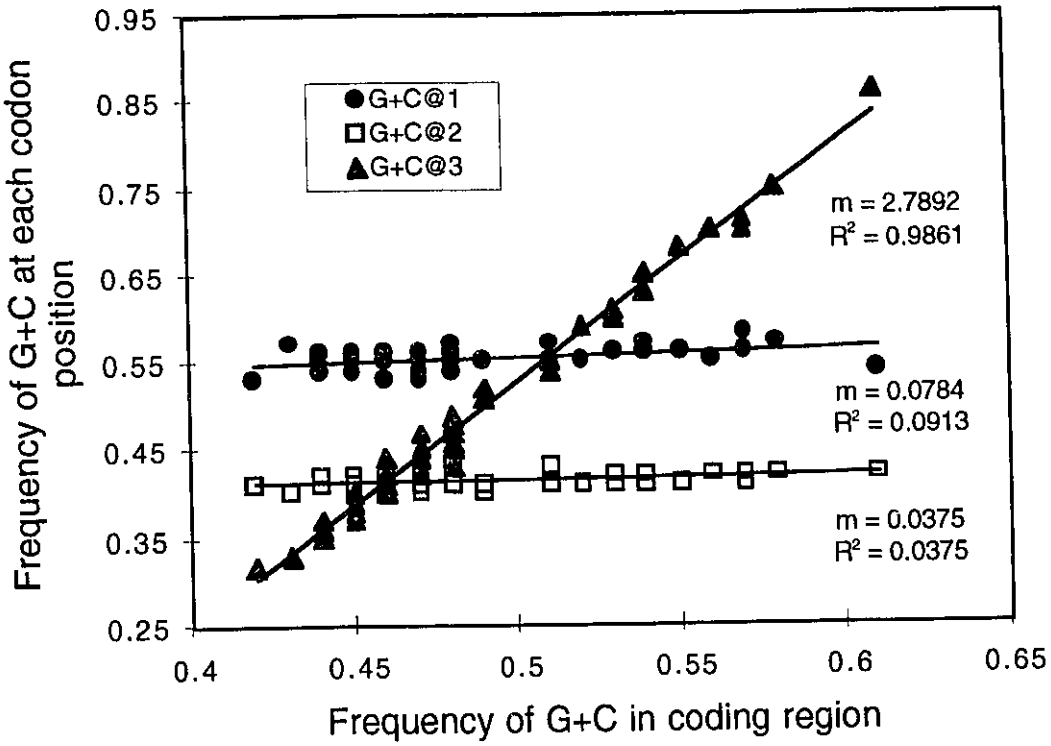

Figure 7: Effect of G+C content of the gene on the G+C content at the first, second and third position of codons.

Effect of intron G+C content on coding region

The G+C content of coding regions is usually characteristic of the overall G+C content of the genome or localized areas where they are located (Mouchiroud and Bernardi, 1993). The G+C content of introns is most likely to represent that of the genome in general since most of it is noncoding and both should therefore converge toward the same G+C content. Plotting the G+C content of the third position of codons from the actin gene versus that of their respective introns shows that the G+C content of the coding region is not significantly influenced by the G+C content of the surrounding non-coding sequences (Figure 8). All plant actins introns studied were found to be AT-rich. Similar observations have been noted for angiosperm introns in general (Goodall and Filipowicz, 1990).

Part III: Phylogeny of actin genes

Phylogeny of angiosperm actin genes

The amino acid and nucleotide alignments were used to generate phylogenetic trees based on neighbor-joining and parsimony methods. Since angiosperm actin genes are known to be monophyletic (Shetterline and Sparrow, 1994; this study, see below), the *Volvox* actin gene is used as the outgroup for all trees generated.

The maximum parsimony and neighbor-joining trees based on the actin amino acid sequences are shown in Figures 9 and 10, respectively. Both these trees are characterized by very low bootstrap values, especially in the deep nodes. The neighbor-joining trees based on the non-synonymous substitutions and the first and second position of codons as well as the maximum parsimony tree based on the first and second position of codons are shown in Figures 11, 12, and 13, respectively. In general, these trees gave higher bootstrap values than trees based on amino acid data. This is likely because the nucleotide data set is larger. The amino acid data set contained

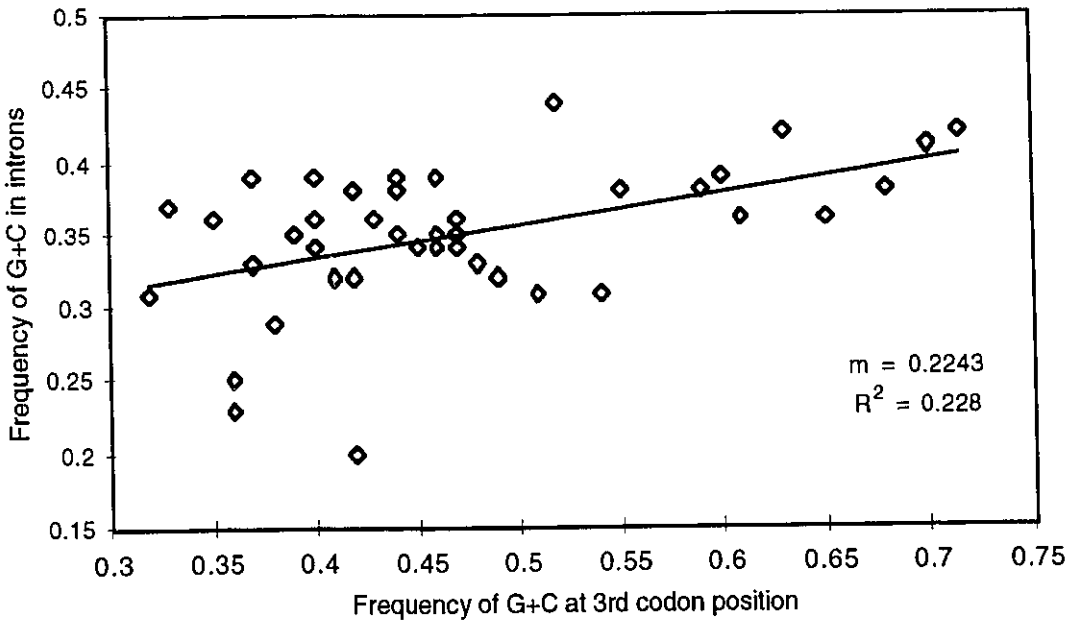

Figure 8: Effect of nucleotide bias in the introns on the codon usage of plant actin genes.

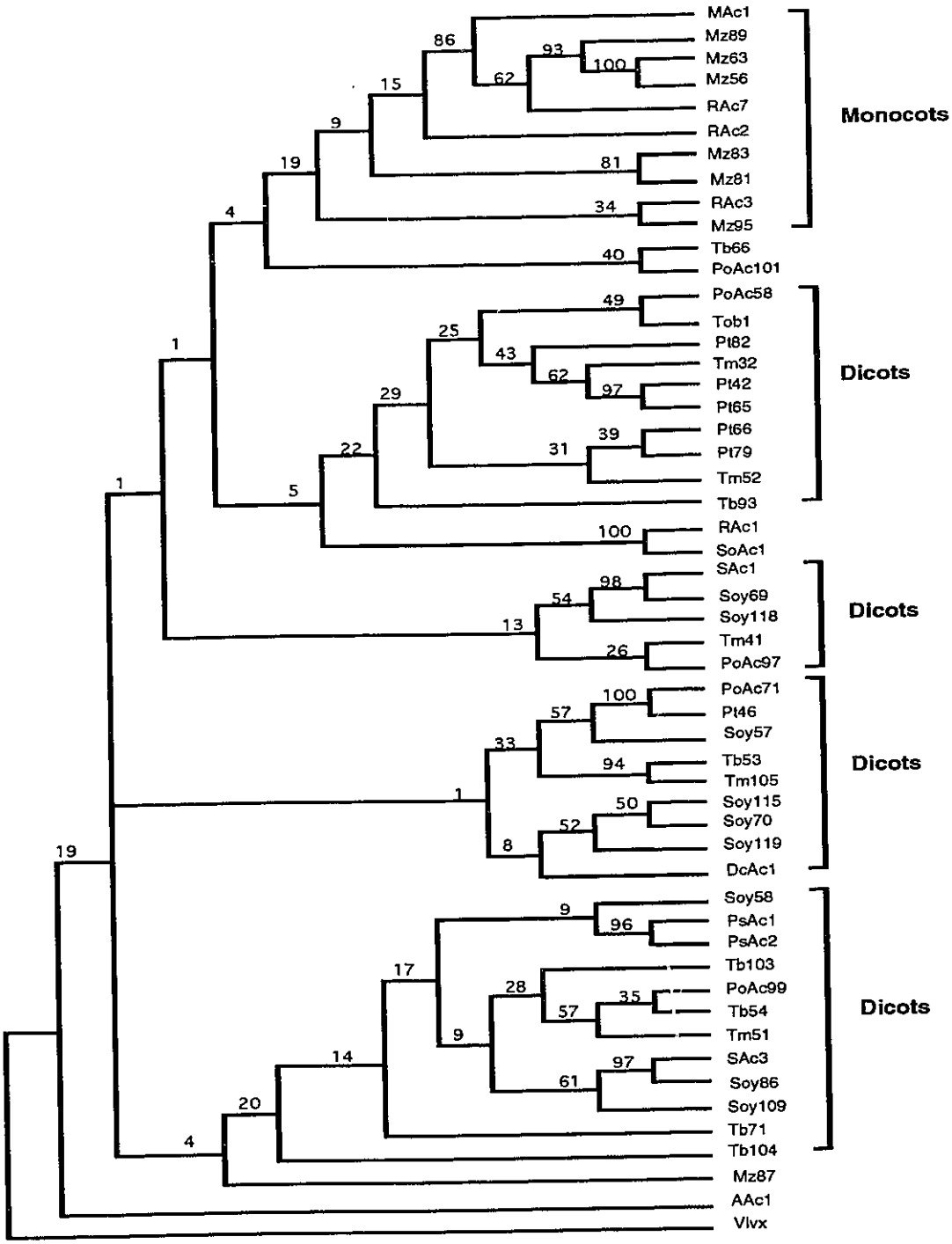

Figure 9: Majority-rule bootstrap consensus tree from parsimony analysis of 53 plant actin amino acid sequences. The *Volvox* actin (Vlvx) is used as the outgroup. Numbers represent the number of times a particular node is present out of 100 bootstrap replicates.

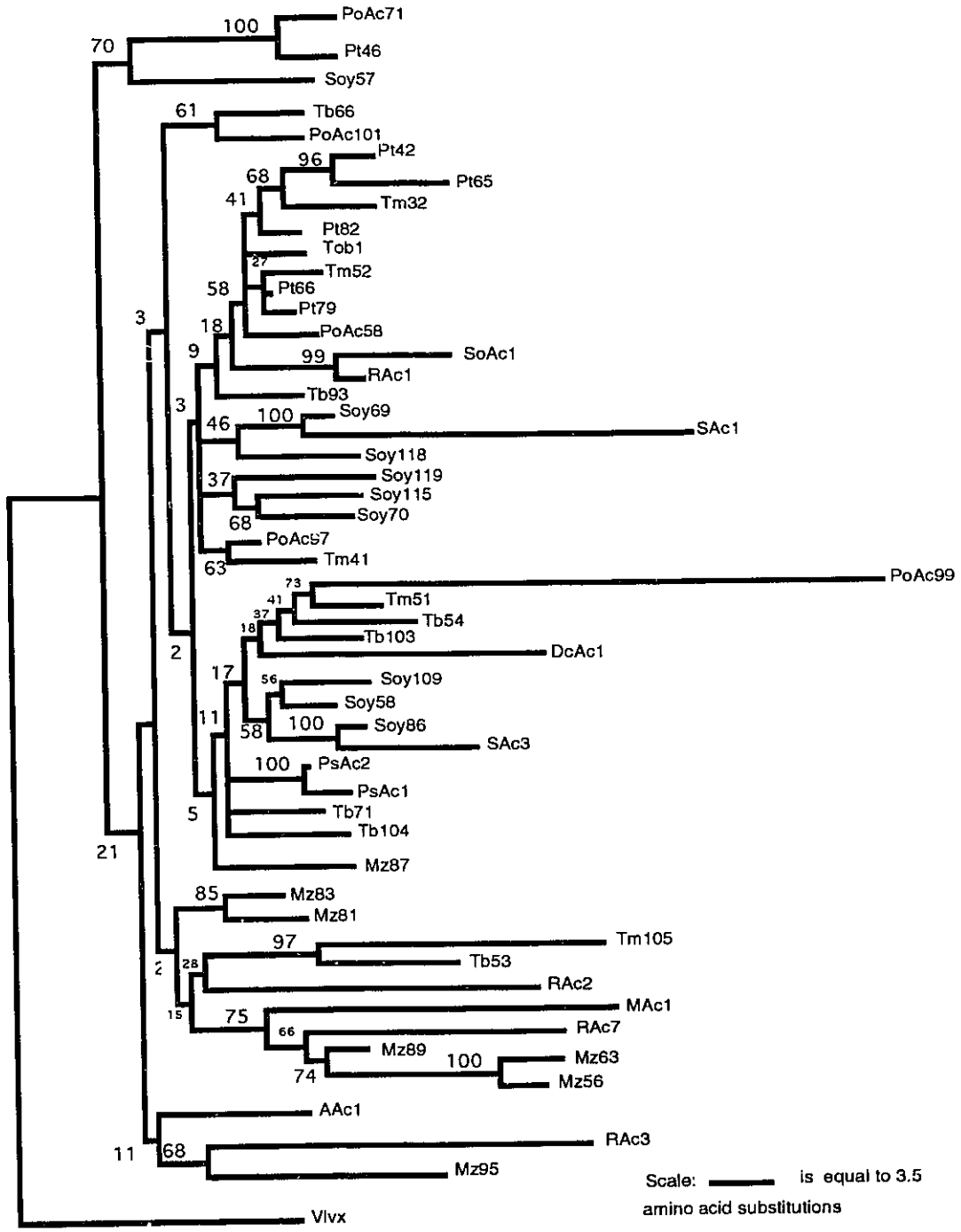

Figure 10: Majority-rule bootstrap consensus tree based on neighbor-joining analysis of 53 plant actin amino acid sequences. The *Volvox* actin (Vlvx) is used as the outgroup. Numbers represent the number of times a particular node is present out of 100 bootstrap replicates. Branch lengths are proportional to the number of amino acid differences (scale is shown).

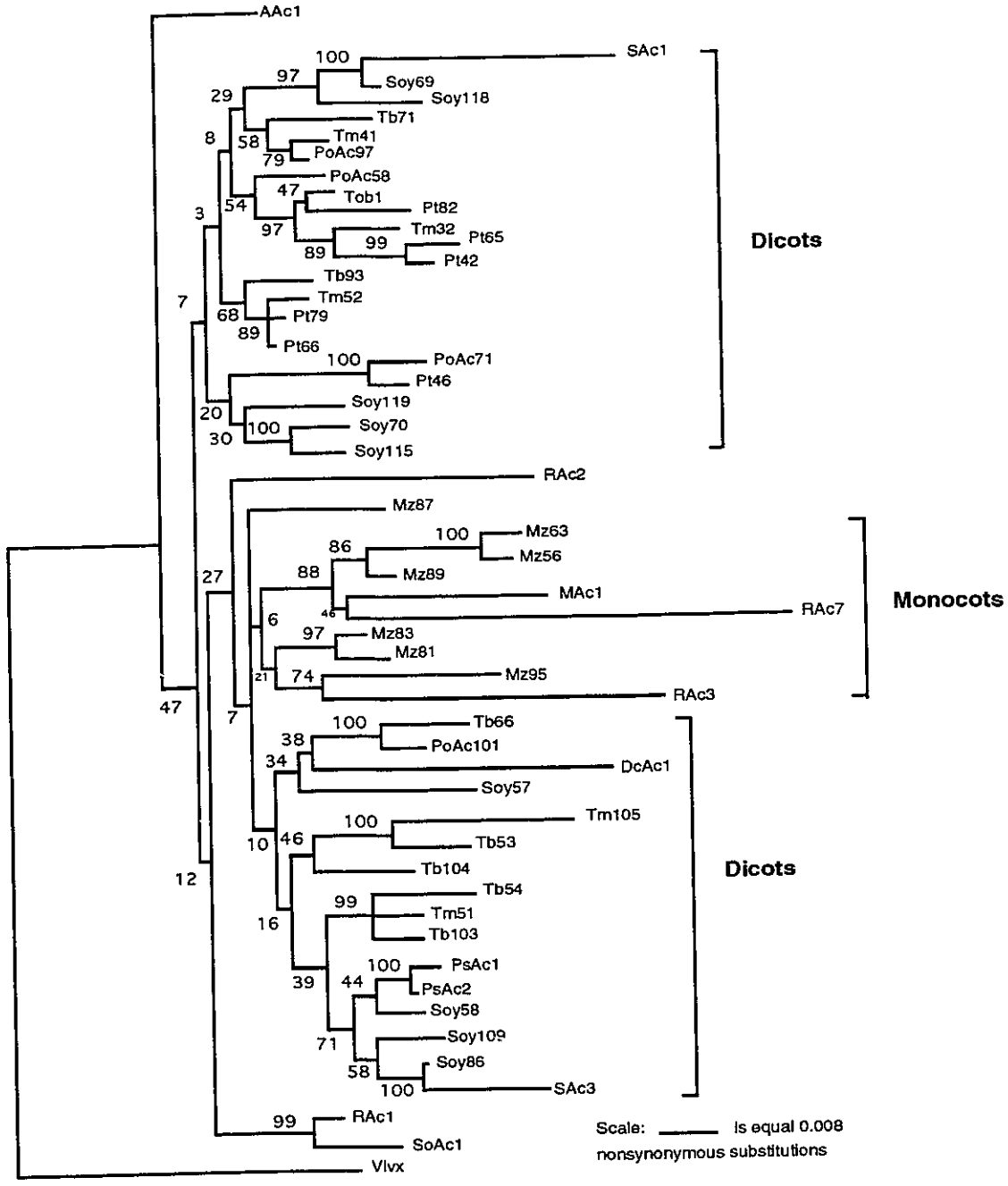

Figure 11: Majority-rule bootstrap consensus tree based of neighbor-joining analysis of Jukes-Cantor corrected proportion of nonsynonymous differences among 52 plant actin nucleotide sequences. PoAc 99 processed pseudogene not included. Numbers represent the number of times a particular node is present out of 100 bootstrap replicates. The *Volvox* actin (Vlvx) is used as the outgroup. Branch lengths are proportional to the inferred proportion of nonsynonymous differences between sequences (scale is shown).

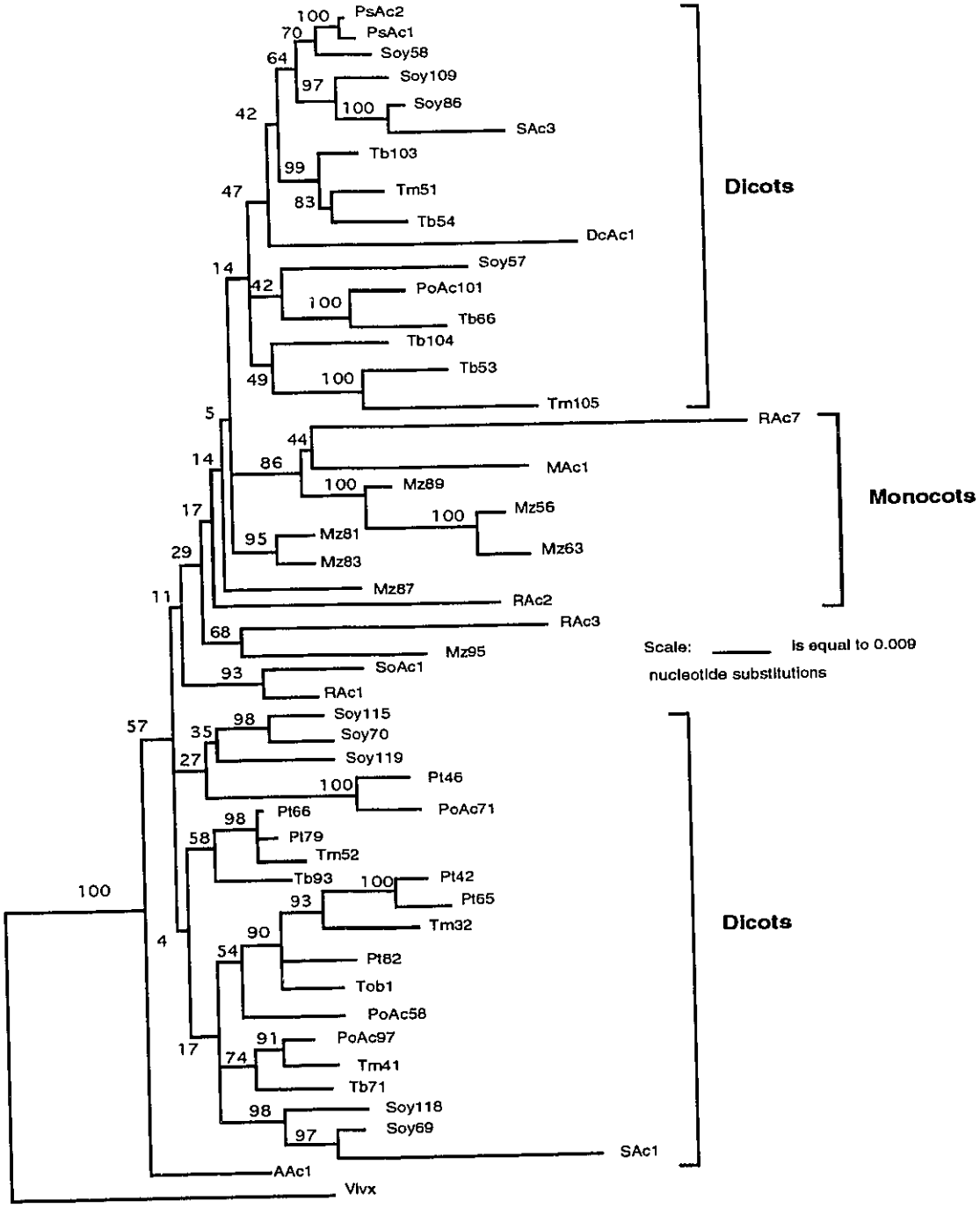

Figure 12 : Majority-rule bootstrap consensus tree based on neighbor-joining analysis of 52 plant actin nucleotide sequences. The *Volvox* actin (Vlvx) is used as the outgroup. PoAc 99 processed pseudogene not included . Numbers represent the number of times a particular node is present out of 100 bootstrap replicates. Branch lengths are proportional to the number of inferred nucleotide substitutions according to the Kimura 2-parameter model (scale is shown). The third site of codons is not included in the analysis.

Figure 13: Majority-rule bootstrap consensus tree from parsimony analysis of 53 plant actin nucleotide sequences. The *Volvox* actin (Vlvx) is used as the outgroup. Third position of codons excluded from analysis. Numbers represent the number of times a particular node is present out of 100 bootstrap replicates. Third position of codons excluded from analysis. Bold lines represent congruent topologies with Neighbor-joining distance trees based on proportion of nonsynonymous differences and first and second position of codons.

a total of 342 sites, 211 of which are variable and 67 of which are parsimony-informative. In contrast, the nucleotide data set for codon positions 1 and 2 consists of a total of 688 sites, 330 of which are variable and 139 of which are parsimony-informative. Thus, the nucleotide data set has about twice as many variable and parsimony-informative sites as the amino acid data set. Therefore, the phylogenetic trees based on the nucleotide data set are probably more reliable since they are based on more information.

In all the plant actin trees (Figures 9-13), the bootstrap values are generally higher for external than internal nodes. Nodes with low bootstrap values indicate that not enough information is available to give firm support for the particular topology. The lack of phylogenetic information supporting these internal nodes suggests that the actin multigene family in plants underwent many duplication events early in its evolution and consequently these internal nodes are clustering together. Because the actin genes are very conserved, not enough changes have occurred to support any particular topology concerning the deep nodes. This is well illustrated in the radial tree shown in Figure 14.

While there is not much support for the arrangement of deep nodes in all the trees, there is nonetheless some congruence between the trees concerning the composition of some of the long branches, especially in the DNA-based trees (Figures 11-14). There appears to be three main lineages of plant actin genes, two of which are comprised of only dicot actin genes and the other consisting of only monocot actin genes arranged paraphyleticaly. However, not all monocot and dicot actin genes fall into these three main lineages. For example, the *Arabidopsis* actin gene, AAc1, consistently comes out as the sister group to all the other angiosperm actin genes (except in the neighbor-joining tree based on amino acid data, for which there is little bootstrap support). Also, the position of the rice and sorghum actin genes, Rac1 and SoAc1, are quite variable with respect to other main branches, as reflected by their low

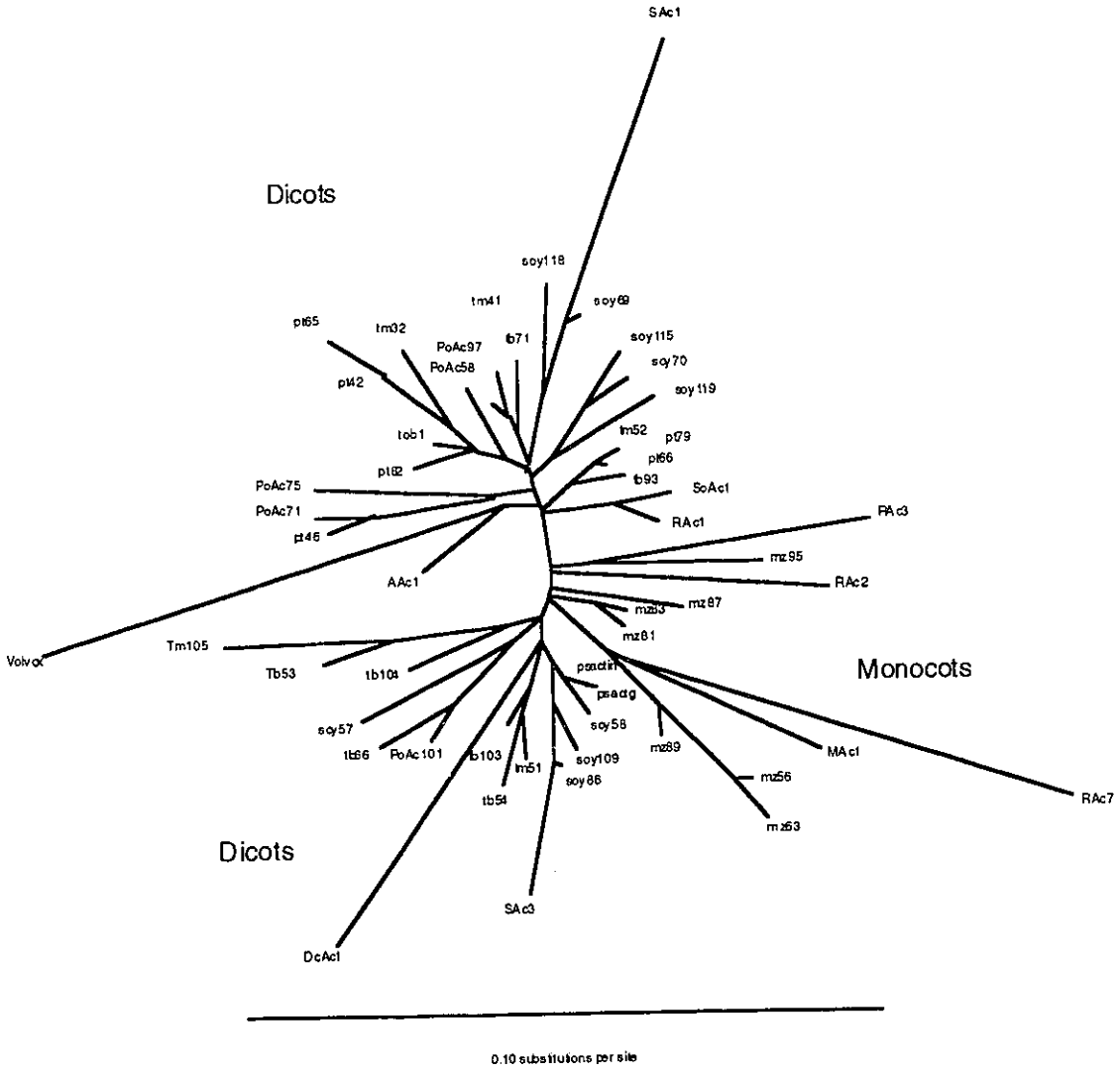

Figure 14: Neighbor-joining radial tree of 53 plant actin nucleotide sequences. The *Volvox* actin gene (Vlvx) is used as the outgroup. Branch lengths are proportional to the number of inferred nucleotide substitutions according to the Kimura 2-parameter model (scale is shown). The third site of codons is not included in the analysis.

bootstrap values. Within each of the three actin lineages, there are genes which always cluster together in all the trees and for which there is good bootstrap support. In Figure 13, the genes which grouped together in all the DNA-based trees have been highlighted in bold. Many of these phylogenetic relationships are also supported by the amino acid-based trees. There is much stability in the topology of the trees within these small clusters of genes. However, the various trees show some variation in the way these clusters of genes are related to one another and this is reflected in their low bootstrap values.

Perhaps the most significant result from the phylogenetic analysis of plant actin genes is that the small consistent clusters of actin genes are composed of genes from more than one species. In addition, the phylogenetic trees show that in several cases the local topologies are similar to the topology of the species tree (Figure 2). Thus, it could be that these genes are related to one another through speciation events and may therefore be orthologous. For example, the DNA-based trees consistently group the genes Pt66, Pt79, Tm52, and Tb93 together and the topology is the same as that for the known relationships of the species to which these genes belong. If this is the case, the two potato genes, Pt66 and Pt79, are the result of a gene duplication which occurred after the speciation event which gave rise to potato and tomato.

In total 6 different groups of potential orthologous genes have been identified. These putative orthologues have been isolated from the larger trees and are shown in Figure 15. One of these groups of potential orthologues is represented by genes from 4 different species, four of these groups are represented by genes from three different species, and one of the groups are represented by genes from only two species. The importance of identifying true orthologous members can not be understated. The use of orthologous genes from species whose divergence times are known is the only way to determine the rate of gene evolution. By identifying different sets of orthologous

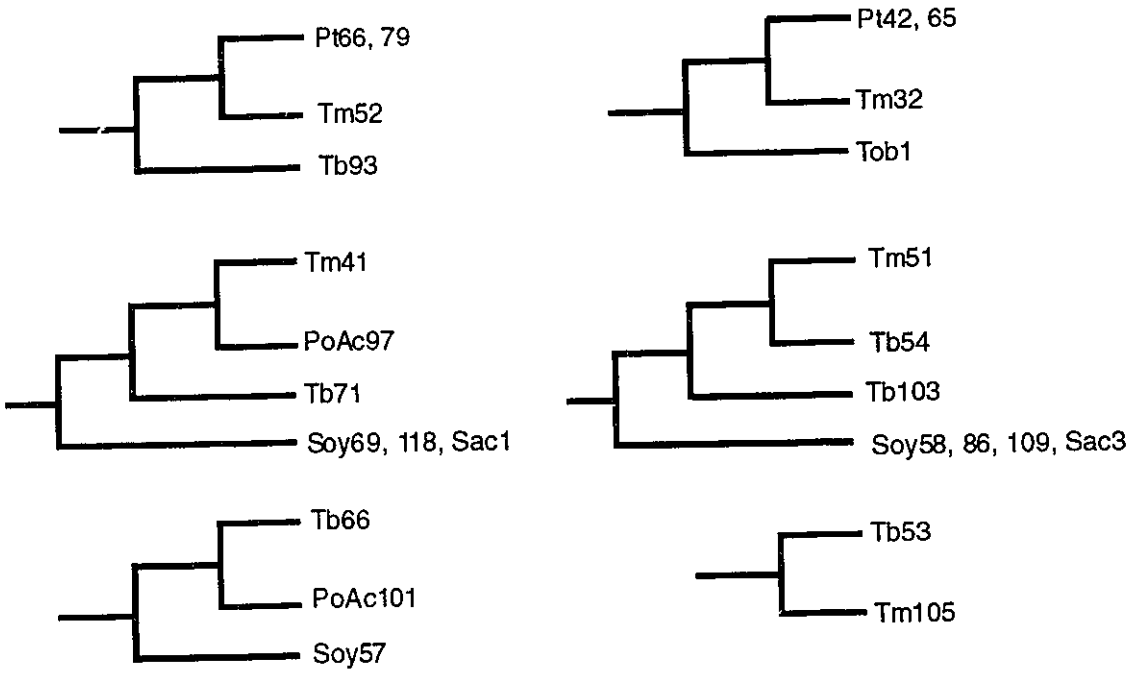

Figure 15 : Putative orthologues identified from phylogenetic trees.

members from different species, independent rates of evolution can be derived and compared. In addition, it is important to calculate rates of evolution using orthologous genes from species with different divergence times. In this way, it is possible to assess if the rate of evolution is constant, i.e., to determine if there is a molecular clock. So far putative orthologous genes have been identified simply by their interrelationships in phylogenetic trees. Other lines of evidence are consistent with the hypothesis that these genes are orthologous. In the case of the potato genes, Pt42 and Pt65, and the tomato gene Tm32, the number of introns possessed by these genes is particularly relevant. As already noted above, Tm32 has lost intron 2 and the two potato genes, Pt42 and Pt65, have lost introns 2 and 3. Given the phylogenetic relationship of these genes, it is possible that intron loss occurred through a two-step process. It could be that intron 2 was lost first in the ancestor of potato and tomato, and intron 3 was lost subsequently in the line leading to potato subsequent to the potato-tomato divergence. Thus, the fact that all three genes, Pt42, Pt65, and Tm32, have lost intron 2 is consistent with their status as putative orthologous members.

The putative orthologue to Tm32, Pt65, and Pt42 in tobacco is Tob1, which has been shown to be expressed only in pollen (Thangavelu et al., 1993). The expression pattern of Tob1 is relevant in this case because the intron loss events in the potato and tomato genes must have occurred in germ line tissue, or tissue leading up to it, in order for these genes lacking introns to be fixed in the species. This suggests that the ancestral gene to Tm32, Pt65, and Pt42, which contained all three introns, was also expressed in the germline tissue and had a similar expression pattern to Tob1. This scenario is consistent with Tob1 being orthologous to Tm32, Pt65, and Pt42.

If orthologous genes belong to species which have diverged relatively recently, it may still be possible to detect some sequence homology between their noncoding sequences, such as introns. The intron sequences of each putative orthologue were

spliced together and used in pairwise comparisons using the sequence search program, SSEARCH, found in the FASTA program package (Pearson, 1994). Splicing the introns together considerably reduced the total number of pair-wise comparisons. Table 7 shows the percent nucleotide sequence identity over a given overlap region as well as the Smith-Waterman scores for the pairwise comparisons of introns from putative orthologues. The Smith-Waterman score is an optimized score which takes into account the total number of nucleotide matches and mismatches, insertions and deletions, and the length of the sequence. The introns of putative orthologues from closely related species, such as potato and tomato (Borisjuk et al., 1993), retain a high level of sequence similarity, about 88% spanning a region of 450-500 nucleotides, and gave high s-w scores. The fact that the percent identity is similar between the introns for both pairs of potato and tomato putative orthologues (ie. Pt66/Tm52 and Pt97/Tm41) is consistent with them having a similar divergence time from a common ancestor. The low s-w score observed between the intron sequences of Pt65 and Tm32, and Pt65 and Pt42, is due to the fact that only intron 1 is present in Pt65 and Pt42, and consequently the region of overlap is much shorter.

The plant lineage leading to soybean and that leading to the Solanaceae last shared a common ancestor about 110 MY ago. Although the fact that angiosperm actin introns all show a relatively high A+T content, and that this might introduce a bias favoring high alignment scores, the aligned regions with more than 55% similarity over more than 150 bases between these two lineages are likely significant (see Figure 8; Pearson and Lipman, 1985). In general, there is a higher sequence identity between the introns of tobacco actins and their putative orthologues in potato and tomato than there is, as expected, between those introns of soybean and those of the putative orthologues in the Solanaceae. Thus, the fact that some similarity still exists between the noncoding

Putative orthologues	Identity of introns per overlap region	s-w score*
Pt66 -Pt79	91.8%/570 nt	1883
Pt 66 - Tm52	89.1%/496 nt	1549
Tb93 - Pt66	61.0%/551 nt	471
Tb93 - Tm52	62.8%/266 nt	279
Pt42 - Pt65	82.7%/75 nt	197
Pt65 - Tm32	79.5%/39 nt	88
Tob1 - Pt65	65.0%/80 nt	85
Tob1 - Tm32	89.7%/29 nt	86
PoAc97 - Tm41	88.3%/454 nt	1382
Tb71 - PoAc97	76.8%/271 nt	606
Tb71 - Tm41	66.4%/434 nt	597
Soy118 - Tm41	55.1%/276 nt	113
Soy118 - PoAc97	57.5%/214 nt	110
Soy118 - Tb71	58.6%/203 nt	132
Soy118 - soy69	57.7%/215 nt	123
Tb66 - PoAc101	72.0%/529 nt	1039
Soy57 - PoAc101	55.7%/221 nt	88
Soy57 - Tb66	57.1%/224 nt	74
Tb54 - Tm51	61.4%/298 n t	250
Tb103 - Tm51	66.6%/329 nt	408
Tb103 - Tb54	62.1%/322 nt	221
Soy58 - Tm51	56.5%/154 nt	97
Soy58 - Tb54	58.1%/155 nt	88
Soy58 - Tb103	59.3%/145 nt	103
Soy58 - Soy86	71.1%/249 nt	407
Soy58 - Soy109	70.0%/277 nt	421
SAc3 - Tm51	59.3%/209 nt	116
SAc3 - Tb54	56.1%/223 nt	118
SAc3 - Tb103	57.5%/179 nt	80
SAc3 - Soy58	69.7%/251 nt	399
SAc3 - Soy86	98.4%/247 n t	936
SAc3 - Soy109	86.3%/255 nt	727

* Smith-Waterman score.
Note: Comparison between Pt65 and Tm32 based on intron 1, and comparison between Tob1 and Tm32 based on introns 1 and 3. Sac1 comparisons not included because this gene is likely a pseudogene (see above).

Table 7: Nucleotide sequence comparisons between introns 1, 2 and 3 of putative plant actin orthologues.

sequence of some tobacco actin genes and those of potato and tomato actin genes is consistent with the hypothesis that these genes are orthologous members.

By assuming a molecular clock or homogeneous rates of substitutions, different sets of orthologous genes from the same species should all have accumulated the same number of nucleotide substitutions since they will all have diverged from a common ancestor at the same time. Thus, by comparing the number of non-synonymous changes, or the synonymous changes if these have not been saturated, between different sets of putative orthologues, it is possible to add support for their status as orthologous members. The independently calculated synonymous and nonsynonymous rates of evolution are not significantly different between different sets of putative orthologues (see below).

One way to test if genes are orthologous is by Southern blot analysis. The idea is that probing a relevant genomic southern using two orthologous members should result in roughly similar patterns of hybridization. Work is currently underway to test orthologues in such a manner.

Phylogeny of eukaryotic actins

In order to put the evolution of plant actin genes in a larger context, representative plant actin amino acid sequences were aligned with those of animals, fungi and protists and used to construct a neighbor-joining tree using the *Giardia lamblia* actin as an outgroup (Figure16). There is very solid ultrastructural, protein-based and rRNA-based phylogenetic studies which suggest that *Giardia* represents one of the most divergent eukaryotes, making it an appropriate outgroup (Gupta et al., 1994; Hashimoto et al., 1994; Cavalier-Smith, 1991b,c; Sogin, 1991).

This tree in Figure 16 supports the idea that the actin multigene family in angiosperms is monophyletic and that their most recent common ancestor is with an

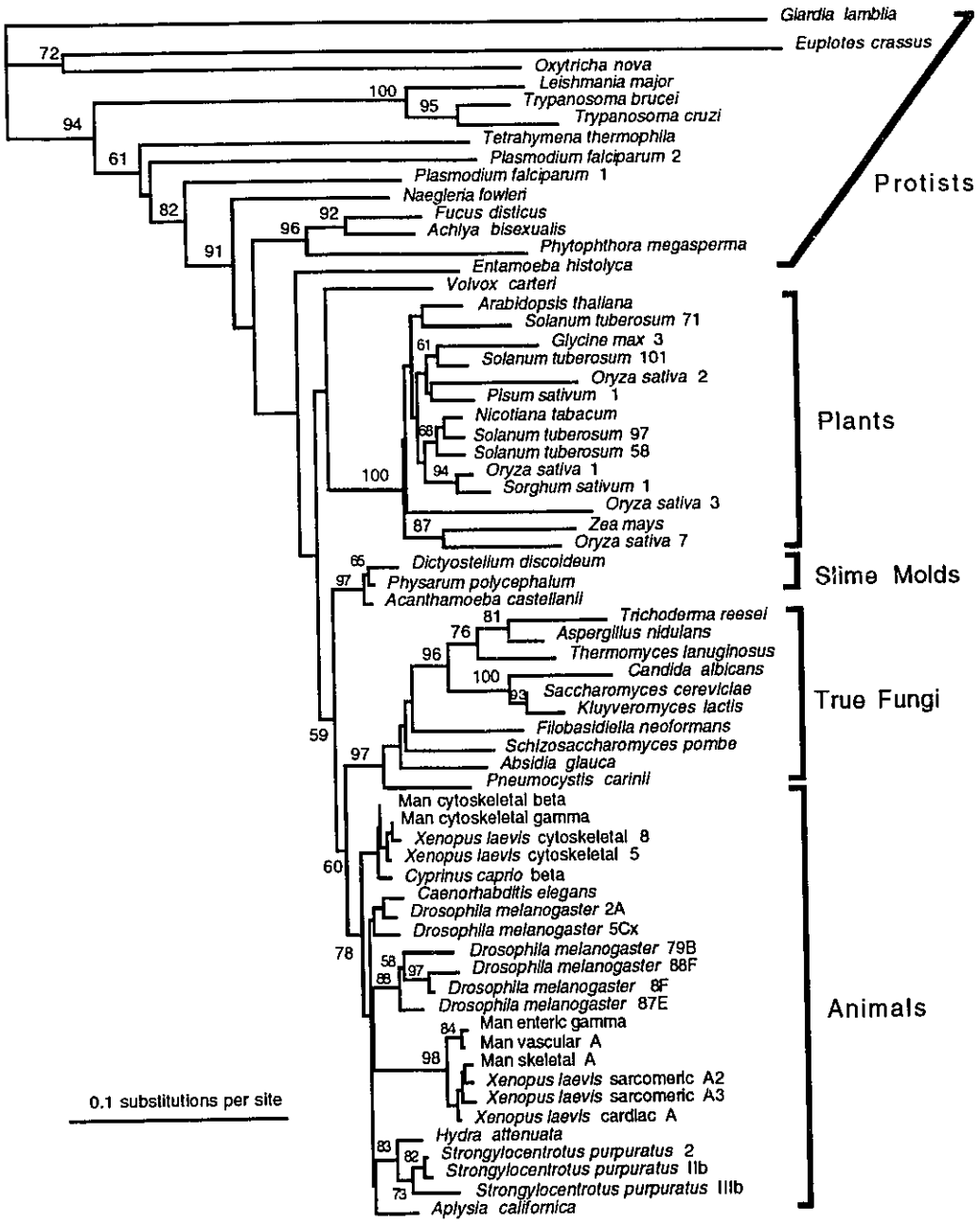

Figure 16: Majority-rule bootstrap consensus tree based on neighbor-joining analysis of 65 eukaryotic actin amino acid sequences. The *Giardia lamblia* actin is used as the outgroup. Numbers represent the number of times a particular node is present out of 100 bootstrap replicates (only values above 50 are shown). Branch lengths are proportional to the number of substitutions per site.

actin gene from green algae. Extrapolating from the gene tree, it is inferred that angiosperms are a monophyletic group. Taking 750 MY as the time of divergence between land plants and green algae (*Volvox*), the tree suggests that plant (angiosperm) actins may all have shared a common ancestor as recently as 300-400 MY ago.

Figure 16 also shows that the rate of evolution of actin genes is not constant among different lineages. Considerably more actin amino acid sequence variation is observed within plants than in animals. This suggests that either the actin multigene family in plants is more ancient than in animals or that it is evolving faster. Humans have last shared a common ancestor with carp and *Xenopus* 400 and 350 MY ago, respectively (Li and Graur, 1991). Yet, judging from the low levels of sequence variation among their cytoplasmic actins, it must be concluded that plant actin genes are, indeed, evolving faster. Similarly, actins in fungi are also evolving faster than in animals. For example, the divergence between *Candida* and *Saccharomyces* is estimated to be about 150 MY ago (Berbee and Taylor, 1993), considerably less than the divergence between carp and humans. Yet the amino acid sequence divergence between these two fungi is substantially greater than that observed between the carp and human actin genes. Since the fungi shown in the tree (Figure 16) all contain single copy actin genes, it is unlikely that the great variation in actin amino acid sequence observed within these fungi is a result of comparing paralogous genes.

The eukaryotic actin tree shown in Figure 16 also shows that actin genes in animals as well as those of fungi are also monophyletic. Thus, the gene duplications which led to the diverse actin isotypes in animals occurred after the appearance of animals. Therefore, one can extrapolate from the actin gene tree to infer that both animals and fungi are monophyletic groups. This conclusion is consistent with phylogenetic trees based on 18S rRNA genes (Cavalier-Smith et al., 1994; Wainwright et al., 1993; Sogin, 1991; Hendriks et al., 1989; Gouy and Li, 1989). The tree also

indicates that animals are more closely related to fungi than either is to plants. In addition, plants, animals and fungi appear to have evolved almost simultaneously, having had a common ancestor relatively recently in terms of eukaryotic evolution.

The slime molds, which are represented by *Dictyostelium discoideum* and *Physarum polycephalum*, also appear to be more closely related to the true fungi and animals than they are to plants. This placement is at odds with the rRNA trees (Cavalier-Smith, 1993; Wainwright et al., 1993; Hendriks et al., 1989; Gunderson et al., 1987). In addition, there is very strong bootstrap support for grouping the slime molds with *Acanthamoeba castellanii.*, a Rhizopod. The actin amino acid tree also provides strong support to the grouping of the brown algae *Fucus disticus* with the Oomycete species *Achlya bisexualis* and *Phytophthora megasperma* and shows that this lineage is not related to higher fungi. These conclusions are consistent with rRNA phylogenies (Cavalier-Smith et al., 1994; Gunderson et al., 1987). *Entamoeba histolyca*, a gut parasite lacking mitochondria, does not group close to *Giardia lamblia* (also lacking mitochondria), but rather comes out as the sister group to plants, fungi, slime molds and animals to the exclusion of other protists. These conclusions are different from rRNA phylogenies (Cavalier-Smith, 1993; Sogin, 1991).

The actin based tree also supports the idea, now becoming widespread, that the deepest branches of eukaryotic evolution lie not within or between plants, fungi and animals, but among the other protist lineages (Schlegel, 1994). This is clearly demonstrated from the relatively long branch lengths, especially those associated with *Giardia*, *Euploites*, *Tetrahymena* and the *Trypanosomes*. This confirms other molecular evidence suggesting that the eukaryotic lineage is a very ancient one, and that perhaps for most of their evolutionary history eukaryotes were amitochondrial unicellular organisms (Schegel, 1994).

Part IV: Gene conversion

The above phylogenetic analysis of 53 plant actin genes shows that in some cases actin genes from the same taxa consistently group together. This could be the result of gene duplication events which occurred after speciation, since the actin multigene family in plants is known to be ancient (see below) and probably underwent gene duplication events prior to the divergence of dicots and monocots. On the other hand, the fact that most of the monocot actin genes consistently group together suggests that gene conversion might also be a factor in the evolution of plant actin genes.

To detect gene conversion among plant actin genes, a computer program was developed (Pratt and Drouin, unpublished) based on the method of Sawyer (1989). Briefly, this program takes a DNA alignment, and compares the distribution of polymorphic informative sites to that of a randomly generated data set.

Using this method, several significant gene conversion events have been detected in the coding regions of maize actin genes. Alignment of both the coding and non-coding regions (introns) of these genes clearly shows the gene regions which have been subject to gene conversion events (Figures 17 and 18). The gene conversion event between the two maize genes, Mz56 and Mz81, is localized to the 3' end of the genes and spans, roughly, the last 130 bases of exon 3 as well as all of intron three (Figure 17). Outside this gene conversion event, nucleotide substitutions are scattered along the coding sequence, and as expected, the intron sequences have diverged to a point where they can no longer be aligned. In contrast, only one nucleotide substitution is observed in the coding region where the gene conversion occurred. Even more striking is that only one nucleotide substitution is observed in intron 4, suggesting that the sequence similarity is definitely due to gene conversion and not to a conservation of sequence due to selective constraints. The lack of nucleotide substitutions within the area of gene

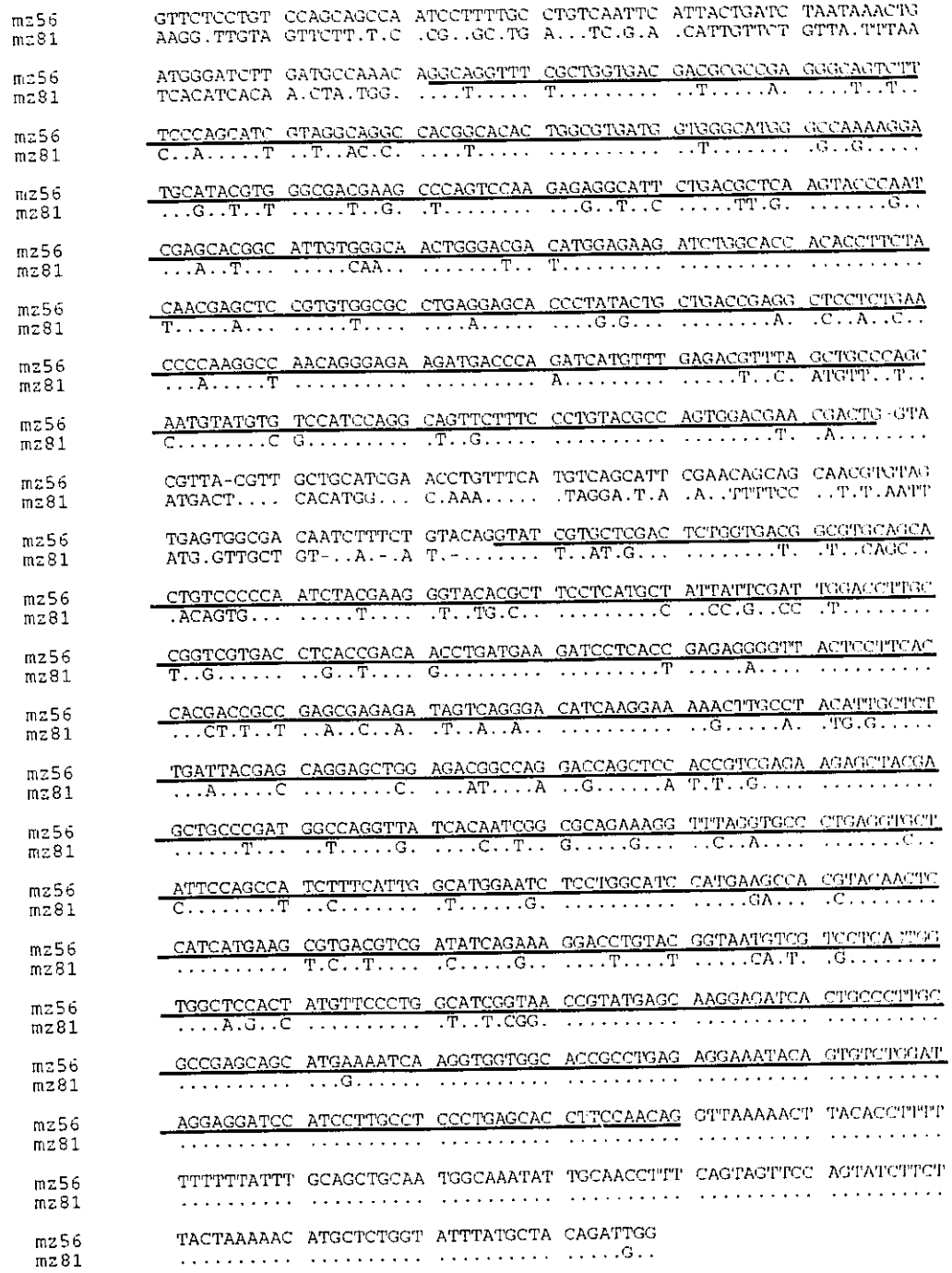

Figure 17: Alignment of two maize actin genes, Mz56 and Mz81, showing gene conversion event in the 3' end of gene. Underlined areas represent exons.

```
mz56 TCCTCTTCCAG CAGCCAATCC TTTTGCCTGT CAATTCATTA CTGATCTAAT AAACTGATGG GATCTTGATG
mz63 .......... .......... .......... .......... .......... .......... ..........

mz56 CCAAACAG-G CAGCTTTCGC TGGTGACGAC GCGCCGAGGG CAGTCTTTCC CAGCATCGTA GGCAGGCCAC
mz63 ...G...... .......... .......... .......... .......... .......... ..........

mz56 GGCACACTGG CGTGATGGTG GGCATGGGCC AAAAGGATGC ATACGTGGGC GACGAAGCCC AGTCCAAGAG
mz63 .......... TC??...... .......... .......... .......... ....G..... ..........

mz56 AGGCATTCTG ACGCTCAAGT ACCCAATCGA GCACGGCATT GTGGGCAACT GGGACGACAT GGAGAAGATC
mz63 .......... .......... .......... .......... .......... .......... ..........

mz56 TGGCACCACA CCTTCTACAA CGAGCTCCGT GTGGCGCCTG AGGAGCACCC TATACTGCTG ACCGAGGCTC
mz63 .......... .......... .......... .......... .......... C......... ..........

mz56 CTCTGAACCC CAAGGCCAAC AGGGAGAAGA TGACCCAGAT CATGTTTGAG ACGTTTAGCT GCCCAGCAAT
mz63 .......... .......... .......... .......... .......... .......... ..........

mz56 GTATGTGTCC ATCCAGGCAG TTCTTTCCCT GTACGCCAGT GGACGAACGA CTG-GTACGT TACGTTGCTG
mz63 .......... .......... .......... .........TG .......... .......... ..........

mz56 CATCGAACCT GTTTCATGTC AGCATTCGAA CAGCAGCAAC GTGTAGTGAG TGGCGACAAT CTTTCTGTAC
mz63 .......... .......... .......... .........-. ......GT.. .......... ........G.

mz56 AGGTATCGTG CTCGACTCTG GTGACGGCGT GCAGCACTGT CCCCCAATCT ACGAAGGGTA CACGCTTCCT
mz63 .......... .......... .......... .......... .......... ....G..... ..........

mz56 CATGCTATTA TTCGATTGGA CCTTGCCGGT CGTGACCTCA CCGACAACCT GATGAAGATC CTCACCGAGA
mz63 .......... .......... .......... .......... .......... .......... ..........

mz56 GGGGTTACTC CTTCACCACG ACCGCCGAGC GAGAGATAGT CAGGGACATC AAGGAAAAAC TTGCCTACAT
mz63 .......... .......... .......... .......... .......... .......... ..........

mz56 TGCTCTTGAT TACGAGCAGG AGCTGGAGAC GGCCAGGACC AGCTCCACCG TCGAGAAGAG CTACGAGCTG
mz63 .......... .......... .......... .......... .......... .......... ..........

mz56 CCCGATGGCC AGGTTATCAC AATCGGCGCA GAAAGGTTTA GGTGCCCTGA GGTGCTATTC CAGCCATCTT
mz63 .......... .......... .......... .......... .......... .......... ..........

mz56 TCATTGGCAT GGAATCTCCT GGCATCCATG AAGCCACGTA CAACTCCATC ATGAAGCGTG ACGTCGATAT
mz63 .......... .......... .......... .......... .......... ......T... ..........

mz56 CAGAAAGGAC CTGTACGGTA ATGTCGTCCT CAGTGGTGGC TCCACTATGT TCCCTGGCAT CGGTAACCGT
mz63 .......... .......... .......... .......... .......... .......... ....G....C

mz56 ATGAGCAAGG AGATCACTGC CCTTGCGCCG AGCAGCATGA AAATCAAGGT GGTGGCACCG CCTGAGAGGA
mz63 .......... .....G.... G.....C..C .......... .GG.G..... ...T..T... ..G.......

mz56 AATACAGTGT CTGGATAGGA GGATCCATCC TTGCCTCCCT GAGCACCTTC CAACAG-GTT AAAAACTTAC
mz63 .G.....C.. T.....T..T ..C.....T. .G.....T.. C..T...... ..G........ GTT.GT..GT

mz56 ACCTTTTTTT TTTATTTGCA GCTGCAATGG CAAATATTGC AACCTTTCAG TAGTTCCAGT ATCTTCTTAC
mz63 TA..GCC.CA AAATC.AAAT T.AA.GT.TC TGCT.GC.TT GCAG..CT.C .GCA.ATGAA G.T.GAA.TA

mz56 TAAAAACATG CTCTGGTATT TATGC
mz63 .T..TCTTGT TAT..TGT.G .G.AG
```

Figure 18: Alignment of two maize actin genes, Mz56 and Mz63, showing gene conversion event. Underlined areas represent exons.

conversion implies that the gene conversion event occurred relatively recently. Since only a relatively small portion of the coding sequence is involved in the gene conversion event, it explains why these two genes do not cluster together in the phylogenetic trees with respect to other monocot genes.

Gene conversion has also detected between Mz56 and Mz63 (Figure 18). This gene conversion event is more extensive than the one described above, also involving Mz56. This gene conversion event spans intron 1 through most of exon 3. Intron 4 is not affected by the gene conversion event and is, as expected, not alignable. The low number of nucleotide substitutions observed in the region of gene conversion indicated that gene conversion occurred relatively recently. Since the gene conversion is large enough to homogenize most of the coding sequences between Mz56 and Mz63, it is not surprising that these two genes cluster together in all the phylogenetic trees. Given the large number of genes in the maize genome, the fact that the maize actin gene Mz56 is involved in both of the gene conversion events detected in this species, might indicate that it is Mz56 that converted the two others.

The fact that the introns were also involved in these two gene conversion events indicate that gene conversion did not proceed via a cDNA intermediate.

Part V: Rates of evolution of angiosperm actin genes

Relative rate tests

In order to determine the rates of evolution of angiosperm actin genes from the known divergence times of orthologous genes and the number of substitutions that have occurred between them, it is important to first determine if the rate of nucleotide substitution is the same for all sets of putative orthologues. If the rate of evolution is constant, then the rate can be applied to all genes and reasonably accurate estimates of their times of divergence can be obtained by simply knowing the amount of nucleotide

substitutions which have taken place between any given pair of genes (assuming no gene conversion).

To determine if the actin genes are evolving at a constant rate, the relative rate test was applied (Li and Tanimura, 1987; Sarich and Wilson, 1973). This test is illustrated in Figure 19. If the rate of evolution has been constant since gene 1 and 2 diverged, the number of nucleotide substitutions between gene 1 and gene 3 and between gene 2 and gene 3 should be the same. The null hypothesis being tested is that the difference in the rate between the two genes (K_{13}-K_{23}) is zero. This test does not require any knowledge of divergence times, but does require that the reference gene (the outgroup) branched off earlier than the divergence between the genes being compared.

Table 8 shows all pair-wise comparisons of the number of non-synonymous differences within each putative orthologous group of genes using the relative rate test. The *Volvox* actin gene (Vlvx) is used as the reference gene for all comparisons. In all cases the difference in rate estimated from K_{13}-K_{23} is not significantly different from zero (Students t-test, $P>0.05$). A power analysis showed that the data is large enough that had there been a significant difference in the rate it would have been detected. Thus, it is concluded that the non-synonymous rate of substitution within each set of putative orthologous genes is constant.

The relative rate test was also applied to the synonymous differences within each putative orthologous group (Table 9). In this case the *Volvox* actin gene could not be used as a reference gene since its time of divergence from plant actin genes is much too great resulting in saturation of substitutions at synonymous sites. Therefore, appropriate outgroups were chosen for each putative orthologous group based on the fact that synonymous differences were not saturated and taking into consideration the results from the phylogenetic analysis. The results (Table 9) indicate that differences in

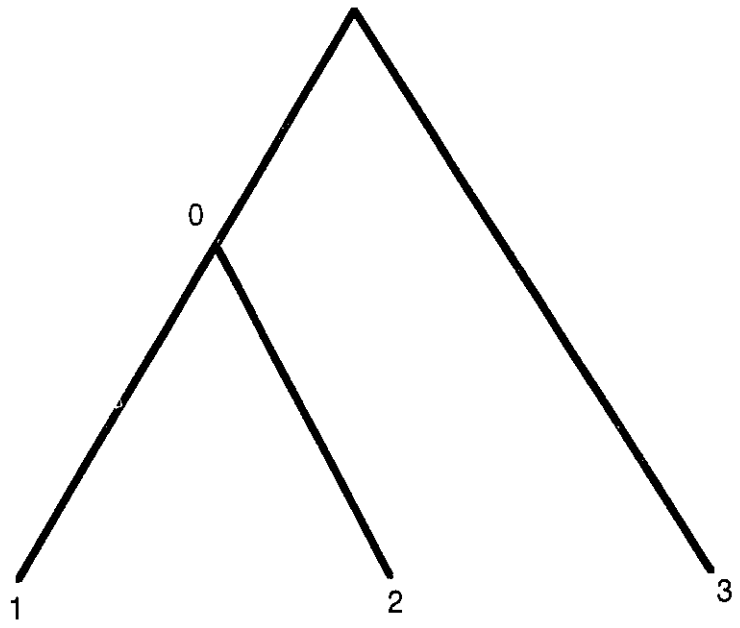

Figure 19: Phylogenetic tree used in the relative rate test. The common ancestor of genes 1 and 2 is denoted by O. The extant outgroup to genes 1 and 2 is denoted by 3.

Gene 1	Gene 2	K_{12}	K_{13}-K_{23}
Tm51	*Tb54*	15.3 ± 4.7	-8.3 ± 15.9
Tm51	*Tb103*	12.4 ± 4.3	-2.7 ± 15.7
Tm51	*Soy58*	15.4 ± 4.8	-3.1 ± 15.6
Tm51	*Soy109*	20.4 ± 5.5	-9.6 ± 16.0
Tb54	*Tb103*	17.5 ± 5.1	5.7 ± 16.1
Tb54	*Soy58*	19.0 ± 5.3	5.2 ± 16.1
Tb54	*Soy109*	26.6 ± 6.3	-1.2 ± 16.4
Tb103	*Soy58*	15.0 ± 4.7	-0.5 ± 15.8
Tb103	*Soy109*	20.9 ± 5.6	-6.9 ± 16.1
Soy58	*Soy109*	13.1 ± 4.4	-6.4 ± 16.2
Tb53	*Tm105*	35.6 ±7.3	-10.7±16.9
PoAc101	*Tb66*	13.5 ± 4.4	-1.7 ± 15.1
PoAc101	*Soy57*	29.9 ± 6.7	-5.5 ± 15.6
Tb66	*Soy57*	31.4 ± 6.9	-3.8 ± 15.6
Tm41	*PoAc97*	5.8 ± 2.9	1.3 ± 16.0
Tm41	*Tb71*	32.8 ± 7.0	-23.5 ± 17.3
Tm41	*Soy118*	22.1 ± 5.7	5.6 ± 15.7
PoAc 97	*Tb71*	30.1 ± 6.7	-24.8 ± 17.2
PoAc97	*Soy118*	19.1 ± 5.3	4.3 ± 16.0
Tb71	*Soy118*	41.8 ± 7.9	29.1 ± 17.0
Tob1	*Pt42*	25.8 ± 6.2	-5.3 ± 15.4
Tob1	*Tm32*	17.6 ± 5.1	-9.7 ± 15.6
Pt42	*Tm32*	25.0 ± 6.1	-4.4 ± 15.9
Tb93	*Pt79*	10.6 ± 3.9	5.8 ± 15.3
Tb93	*Tm52*	15.0 ± 4.7	2.5 ± 15.5
Tm52	*Pt79*	4.35 ± 2.5	3.2 ± 15.1

The average number of non-synonymous sites compared (L) is 688.
None of the relative rate differences (K_{13}-K_{23}) are significantly different from zero at the 5% level.
A positive value means that gene 1 evolves faster, a negative value means that gene 2 evolves faster.

Table 8: Relative rate test using the number of non-synonymous substitutions per 1000 sites. *Volvox* actin gene is used as the outgroup. Number after the ± sign is the standard error.

Gene 1	Gene 2	K_{12}	K_{13}-K_{23}
Outgroup = MAc1			
Tm51	*Tb54*	57.1 ± 7.5	-11.7 ±40.1
Outgroup = Soy57			
PoAc101	*Tb66*	35.1 ± 5.2	10.2 ± 30.5
Outgroup = Soy118			
Tm41	*PoAc97*	7.9 ± 2.2	14.7 ± 21.6
Tm41	*Tb71*	28.3 ± 4.5	1.2 ± 23.2
PoAc 97	*Tb71*	25.9 ± 4.3	-13.5 ± 21.8
Outgroup = Tb104			
Tb53	*Tm105*	29.9±4.7	-1.2±10.5
Outgroup = PoAc58			
Tob1	*Pt42*	60.1 ± 7.7	11.1 ± 26.2
Tob1	*Tm32*	86.5 ± 10.9	-24.7 ± 31.4
Pt42	*Tm32*	50.6 ± 6.9	-36.0 ± 30.7
Tb93	*Pt79*	32.8 ± 5.0	34.0 ± 37.4
Tb93	*Tm52*	32.6± 5.0	27.4 ± 38.4
Tm52	*Pt79*	9.2 ± 2.4	6.1 ± 32.0

The average number of synonymous sites compared (L) is 184.
None of the relative rate differences (K_{13}-K_{23}) are significantly different from zero at the 5% level.

Table 9: Relative rate test using the number of synonymous substitutions per 100 sites. A constant outgroup could not be used for all genes because of saturation at synonymous sites between divergent genes. Number after the ± sign is the standard error.

the synonymous rates are also not significantly different from zero (P>0.05). Thus, the genes within each putative orthologous set are evolving at the same rate at their synonymous sites.

The above relative rate tests were confined to testing for differences in rate within groups of putative orthologous genes. To determine if the rates of evolution in actin genes differ between species or between monocots and dicots, the mean number of non-synonymous differences for actin genes within each species was calculated relative to the *Volvox* actin gene (Table 10) and used in the relative rate test (Table 11). The results indicate that actin genes in almost all the lineages studies here are evolving at a roughly similar rate. A power analysis showed that the data set is large enough that had there been a significant difference in the rate it would have been detected. However, the rice actin genes appear to be evolving significantly faster than those of potato at an α level of 5% but not at the Bonferroni corrected α level of 0.33%. It is noted however, that, although not significally different from zero, the actin genes of monocots are evolving slightly faster than those of dicots. It is possible that had a closer outgroup been used (rather than the *Volvox* actin gene), statistically significant differences might have been revealed.

To determine if actin genes are evolving at similar rates among different eukaryotic lineages, the mean number of estimated nucleotide substitutions at the second position of codons was calculated relative to the *Fucus disticus* actin and used in the relative rate test (Table 12). The second position of codons was used because, in eukaryotes as a whole, a correlation was found between the first position of codons and the G+C content of the gene (R^2=0.5). The results indicate that the differences in the rate of evolution of actin genes between animals, fungi, and angiosperms are not sig. icantly different. The actin genes in slime molds are evolving significantly slower

Taxa	K_n
Tomato	81.17 ± 11.35
Potato	75.23 ± 10.88
Tobacco	77.97 ± 11.10
Soybean	83.49 ± 11.54
Maize	87.65 ±11.87
Rice	110.32 ± 13.48
Dicots	80.13 ± 11.27
Monocots	94.20 ± 12.36

Table 10: Mean number of non-synonymous (Kn) substitutions and standard errors (per 1000 sites) for all actins within each taxa relative to the *Volvox* actin. The total number of non-synonymous sites is 688. The mean for all dicot species include data from all species shown in Figure 11. The mean for monocots include all the monocot species shown in Figure 11. (The following genes were ommited from analysis because they are pseudogenes: Tm32. Pt42, Pt65, Tb71, Sac1 and Mz65).

TAXON 1	TAXON 2	K_{13}-K_{23}
Tomato	Potato	5.94 ± 15.72
Tomato	Tobacco	3.20 ± 15.86
Tomato	Soybean	-2.32 ± 16.19
Tomato	Maize	-6.48 ± 16.42
Tomato	Rice	-29.15 ± 17.62
Potato	Tobacco	-2.74 ± 15.54
Potato	Soybean	-8.26 ± 15.86
Potato	Maize	-12.42 ± 16.10
Potato	Rice	* -35.09 ± 17.32
Tobacco	Soybean	-5.52 ± 16.01
Tobacco	Maize	-9.68 ± 16.25
Tobacco	Rice	-32.35 ± 17.46
Soybean	Maize	-4.16 ± 16.56
Soybean	Rice	-26.83 ± 17.74
Maize	Rice	-22.67 ± 17.96
Dicot	Monocots	-14.07 ± 16.73

* Significant at 5% level but not at the Bonferroni corrected α level of 0.33%.

Table 11: Relative rate test based on the mean number of non-synonymous substitutions (per 1000 sites) for all actins within each taxa relative to the outgroup, *Volvox* actin. The number after the ± sign is the standard error of the relative rate difference. The mean for all dicot species include data from all species shown in Figure 11. The mean for monocots include all the monocot species shown in Figure 11.

TAXON 1	TAXON 2	K_{13}-K_{23}
Animals	Angiosperms	-2.2 ± 2.63
Animals	Fungi	0.16 ± 2.46
Animals	Slime Molds	3.62 ± 2.65
Angiosperms	Fungi	2.36 ± 2.61
Angiosperms	Slime Molds	* 5.82 ± 2.39
Fungi	Slime Molds	3.46 ± 2.20

* Significant at the 5% level. However, it is not significant at the Bonferroni corrected α level of 0.8%.

Table 12: Relative rate test based on the mean number of nucleotide substitutions at the second position of codons (per 100 sites) for all actins within each taxa relative to the outgroup, *Fucus disticus* actin. Nucleotide substitutions were calculated using the Kimura 2-parameter model. The number after the ± sign is the standard error of the relative rate difference. Actin genes within each taxa are those shown in Figure 16.

than those of angiosperms at an α level of 5%, but not at the Bonferroni corrected α level of 0.8%. These results suggest that actins in all of these lineages are evolving at the same rate. However, it must be kept in mind that the number of nucleotide substitutions accumulated have been averaged along a very long time span. The evidence presented below suggests that within a more recent evolutionary time span, the rate of evolution of actin gene varies among different eukaryotic lineages.

Rates of evolution

By definition, the divergence time between two orthologous genes is equal to twice the divergence time between the two species which they come from. Since the divergence time between tobacco and potato/tomato is known (40 MY), as well as the divergence time between the Solanaceae and the lineage leading to soybean (110 MY), the rate of evolution can be calculated for orthologous actin genes from these species.

Table 13 shows the number of synonymous (K_s) and non-synonymous (K_n) differences between putative orthologues as well the corresponding rates of synonymous and non-synonymous substitutions based on the appropriate times of divergence. In general, the 18 independent estimates of the rate of synonymous and non-synonymous substitutions are very similar. This is consistent with the idea that the genes used to calculate the rates really are orthologues and that dicot actins are evolving at the same rate. There is no significant difference between the rates of synonymous and non-synonymous substitutions calculated using the 40 and 110 million year (MY) divergence times. Average rates (from all 18 independent calculations) of synonymous and non-synonymous substitutions for dicot actin genes are 5.44×10^{-9}/site/year and 0.13×10^{-9}/site/year, respectively. The synonymous substitution rate for dicot actin

Putative orthologues	K_n	Rate K_n	K_s	Rate K_s
40 million years				
Tb93-Tm52	1.50± .47	.19± .02	32.64±4.98	4.1± .62
Tb93-Pt66	1.06± .40	.13± .05	30.34±4.71	3.8± .59
Tb93-Pt79	1.06± .40	.13± .05	32.77±5.03	4.1± .63
Tob1-Tm32	1.76± .51	.22± .06	86.49± 10.97	10.8± 1.4
Tob1-Pt42	2.12± .56	.26± .07	60.05±7.72	7.5± .96
Tob1-Pt65	2.58± .62	.32± .08	62.71±7.97	7.8± 1.0
Tb71-Tm41	3.28± .70	.41± .09	28.29±4.53	3.5± .57
Tb71-PoAc97	3.01± .67	.38± .08	25.94±4.26	3.2± .53
Tb54-Tm51	1.53± .47	.19± .06	57.06± 7.48	7.1± .93
Tb66-PoAc101	1.35± .44	.17± .05	35.14± 5.18	4.4± .65
Tb53-Tm105	3.56±.73	.44±.09	29.91±4.72	3.74±59
Average		*0.21(0.12)*		*4.54 (1.28)*
110 million years				
Soy69-Tm41	2.06± .55	.09± .02	139.71± 19.17	6.4± .87
Soy69-PoAc97	1.96± .51	.09± .02	129.63± 17.13	5.9± .78
Soy69-Tb71	4.62± .84	.21± .04	136.38± 22.65	6.2± 1.0
Soy118-Tm41	2.21± .57	.10± .02	122.81± 16.33	5.6± .74
Soy118-PoAc97	1.91± .53	.09± .02	108.11± 14.18	4.9± .64
Soy118-Tb71	4.18± .79	.19± .04	121.63± 16.50	5.5± .75
Soy57-Tb66	3.14± .68	.14± .03	135.45± 20.22	6.2± .92
Soy57-PoAc101	2.99± .67	.14± .03	145.57± 22.77	6.6± 1.0
Soy58-Tb103	1.50± .47	.07± .02	131.68± 18.95	6.0± .86
Soy58-Tb54	1.90± .53	.09± .02	143.40± 20.39	6.5± .93
Soy58-Tm51	1.53± .47	.07± .02	126.19± 17.49	5.7± .79
Soy109-Tb103	2.09± .56	.09± .02	121.16± 16.65	5.5± .76
Soy109-Tb54	2.66± .63	.12±.03	123.29± 16.19	5.6± .74
Soy109-Tm51	2.04± .55	.09± .02	124.75± 17.35	5.7± .79
Average		*0.10 (0.02)*		*5.88 (.49)*
Total Average		***0.13 (0.08)***		***5.44 (1.03)***

Note: Averages do not include rates calculated using Tb71, Pt42, Pt65, and Tm32, since these are presumably nonfunctional

Table 13: Number of synonymous (K_s) and nonsynonymous (K_n) substitutions per 100 sites between putative orthologues. Rates of evolution are calculated according to twice the time since the species had a common ancestor and are shown in units of $x10^{-9}$/site/year. The number after the ± is the standard error. Number in parenthesis is the standard deviation.

genes agrees well with that for other plant nuclear genes (Wolfe et al., 1987) and other eukaryotic nuclear genes (Li and Graur, 1991).

The rate of non-synonymous substitutions of dicot actin genes is relatively low, suggestive of strong functional constraints operating on the protein. This is consistent with the suggestion that actin function is highly conserved in all eukaryotes. However, the rate of non-synonymous substitution in dicot actin genes appears to be 4-10 times higher than in mammalian alpha and beta actin genes. In fact, Li and Graur (1991) calculated, using rodent and primate genes, the rate of non-synonymous substitution for alpha and beta actin genes to be $0.01 \pm 0.01 \times 10^{-9}$/site/year and $0.03 \pm 0.02 \times 10^{-9}$/site/year, respectively. Although it is possible that the divergence times used to calculate the rate of evolution of dicot actin genes in this study have been underestimated, to assume that the rates are equal in plant and mammalian actins would imply that the tobacco diverged from the lineage leading to potato and tomato over 300 MY ago. This date is prior to that thought for the origin of angiosperms (Stewart, 1983).

The average rates of synonymous and non-synonymous substitutions were also calculated from single copy actins from a number of fungi whose times of divergence have been established (Table 14)(Berbee and Taylor, 1993). The average non-synonymous rate of substitution was found to be 0.13×10^{-9}/site/year, virtually identical to that in plants. The average synonymous rate appears to be a little lower than that of plant actin genes. However, this could be the result of saturation at synonymous sites.

Genes	K_n	Rate K_n	K_s	Rate K_s
75 million years				
Kluyve-Saccharo	2.00 ± .48	.13 ± .03	54.19 ± 7.33	3.61 ± .48
135 million years				
Kluyve-Candida	4.34 ± .72	.16 ± .03	70.64 ± 8.52	2.62 ± .32
Saccharo-Candida	3.90 ± .68	.14 ± .03	66.86 ± 8.27	2.48 ± .31
335 million years				
Kluyve-Schizo	7.71 ± .97	.12 ± .01	saturated	NA
Saccharo-Schizo	6.17 ± .86	.09 ± .01	saturated	NA
Candida-Schizo	8.69 ± 1.04	.13 ± .02	saturated	NA
Average		0.13 (0.02)		2.90 (0.62)

Table 14: Number of synonymous (K_s) and non-synonymous (K_n) substitutions per 100 sites between single-copy actin genes of four fungal species. Rates of evolution are calculated as in legend to Table 13.

Discussion

Actin gene copy number

Based on the number of actin genes cloned from potato and soybean in this study and the number of matches with those already present in the Genbank/Embl data base, the number of actin genes in these species is estimated to be about 20-30 copies per genome. Although the method used to generate this estimate is susceptible to a large degree of error because it requires many assumptions, which may not have been met, nonetheless, our estimates are in good agreement with previously published results. Southern blot analysis of tobacco, potato, and soybean genomic DNA using the soybean Sac3 as a probe revealed the presence of at least 25-30 genes per genome (Thangavelu et al., 1993). Bernatzky and Tanksley (1986a) have shown using southern analysis that there may be as many as 40 actin genes in tomato, and in petunia the actin multigene family is thought to be comprised of 100-200 members belonging to six diverse subfamilies (McLean et al., 1988 and 1990). The *Arabidopsis* genome may contain at least 12-20 diverse gene members (McDowell et al., unpublished data, referenced in Meagher, 1991). Based on the fact that none of the actin sequences identified in this study matched the previously cloned Mac1, it is concluded that the actin multigene family in maize is also relatively large.

The fact that all these dicots may have the relatively same number of actin genes in their genomes is interesting since their genome sizes vary considerably. For example tomato has a genome size more than five times smaller than that of tobacco and maize (Tanksley, 1987). *Arabidopsis* has one of the smallest plant genomes, 20 times smaller than that of soybean (Arumuganathan and Earle, 1991). In addition, since tomato is a diploid (Tanksley, 1987) and the cultivated potato (*Solanum tuberosum*), a very close relative (Borisjuk et al., 1993), is a tetraploid (Hosaka, 1986), it would be natural to

expect that potato would have twice the number of actin genes (assuming potato is not an allotetraploid). The conservation of actin gene copy number across different plant species with varying genome size and ploidy levels suggests a functional role for the various actin members. It also suggests that during polyploidization the number of actin loci or genes does not increase. Such a lack of additivity of parental genes in polyploid species has previously been reported (reviewed in Soltis et al., 1992).

Genetic linkage analysis has shown that in tomato actin genes map to as many as 10 different loci (Bernatzky and Tanksley, 1986a). It is likely that potato has a similar gene organization since their genomes are very similar (Helentjaris, 1993; Borisjuk et al., 1993). In *Petunia hybrida* the six diverse actin gene subfamilies map to five locations on five different chromosomes (McLean et al., 1988; 1990). All of these loci contain multiple copies of actin which are tightly linked and are very similar to one another. These results suggest that actin genes in different loci may represent ancient gene duplication events, whereas the clusters of genes within each loci originated more recently through unequal crossing-over events (McLean et al., 1990). The similarity of actin genes within the same cluster could also mean that these genes are undergoing concerted evolution through unequal crossing over events and/or gene conversion. If this is the case, then it would explain why not as many actin genes were cloned as have been estimated for each of the five species in this study. If this is the case, then most of the major lineages of actins have probably been cloned. The fact actin genes from each species are represented in most of the major branches of the phylogenetic trees is consistent with this idea.

The large number of actin genes present in plants is in sharp contrast to that observed in other eukaryotic lineages. In fungi and protists actins are found as either single copies or as part of a small multigene family (reviewed in Sheterline and Sparrow, 1994). In *Drosophila melanogaster* and *C. elegans* the actin multigene family

is comprised of six and four members, respectively (Vandekerckhove and Weber, 1984; Beifuss and Durica, 1992). In higher vertebrates, genetic analysis of the actin multigene family has shown that there are 20-30 actin genes within the genome, but only about six are functional and the rest appear to be processed pseudogenes (McHugh et al., 1991). The actin multigene family in animals has been shown to be developmentally regulated in a temporal and tissue specific manner (Beach and Jeffery, 1992; McHugh et al., 1991). Although it has been shown that the different subfamilies of actin genes in plants are also differentially regulated (reviewed in Reece et al., 1992), it is not apparent why they would need so many actin genes. Since actin is not present in large amounts in any tissue, it is unlikely that the large number of genes is required to direct the synthesis of enough protein (Meagher, 1991). The presence of this large and diverse family in plants could reflect a more diverse role for actins than has been previously established in animals. Alternatively, there could be some mechanism in plants responsible for actin gene amplification which is not present or is not as efficient in other eukaryotes. Plant genomes are much more tolerant to genome size increases than animals as exemplified by the observation that polyploidy is much more common in plant species that animal species (Soltis et al., 1992), and the observation that the amount of DNA per haploid genome (C value) can vary substantially between plant species (Arumuganathan and Earle, 1991). More fine tuned studies at the cellular level need to be done to determine more precisely the role of this large and diverse gene family.

Conservation of number and location of angiosperm actin introns

All angiosperm actin sequences contain three introns at identical locations along the amino acid sequence. The first intron position of angiosperm actin genes is shared

with *Volvox carteri* and *C. elegans*. The second intron position is shared with two *Plasmodium* actin genes as well as a sea star actin gene and muscle actin genes in vertebrates. The last intron position is shared with both actin genes from the nematode, *Onchocerca. volvulus.* Why these three intron positions are so conserved in plant actins but not in other lineages is not understood. This study almost tripled the number of actin genes available for plants and yet no new intron locations were identified. Since the number and location of introns in angiosperms is unique among eukaryotes, it suggests that all angiosperm actin genes are derived from the same common ancestor. The monophyly of angiosperm actin genes is also supported by the presence of unique amino acids shared between all angiosperm actins at their amino terminal (Reece et al., 1992), by a number of phylogenetic studies (Sheterline and Sparrow, 1994; Baldauf and Palmer, 1993; Reece et al, 1992; Hightower and Meagher, 1986), as well as this study.

It has been suggested that the ancestral metazoan actin gene(s) included all the introns found in more highly evolved forms, presumably demarcating functional domains, and that, in the various branches of evolution, these genes underwent selective deletion of different introns (Reece et al., 1992; Holland and Blake, 1990; Wildeman, 1988). A contrasting view states that the pattern of intron locations could be explained by the insertion of new introns coupled with the low-level multiplication of these genes at various stages of evolution (Dibb and Newman, 1989). According to this model, it is extremely unlikely that all 40 known positions of introns demarcate the primitive functional domains of the protein, or its evolutionary building blocks. It is implied, instead, that introns may behave like transposons, inserting and deleting precisely at their termini. In addition, it is argued that there are protosplice sequences which promote the acquisition of a new introns, and because of the conserved nature of actin sequences it assures that introns arise in the same sites over and over again (Dibb

and Newman, 1989). However, angiosperm actins evolve faster than vertebrate actins and it is in angiosperm actins that intron position is most conserved.

It is interesting to speculate about whether intron loss and gain in plants is generally less frequent than in other lineages. In a recent review, Reece et al. (1992) addressed the question of whether the rate of intron insertion and/or deletion was lower in general for plants than in animals or whether this phenomenon was restricted to plant actin genes only. It was concluded that intron and/or deletion is indeed less frequent in relation to the rate of sequence divergence for plants than for fungi or animals. However, more homologous genes need to be compared between plants and other eukaryotic lineages to strengthen this conclusion. An adequate explanation for the conservation of intron positions in plant actin genes is still lacking.

Since actin is not found in prokaryotes it is not considered to be an 'old' protein. Studies on the arrangement of introns in actins are relevant to the 'introns early' versus 'introns late' debate. According to the 'introns early' theory, modern genes are the result of the assembly of exons encoding folding regions, structural elements and functional domains via intron-mediated recombinations (Gilbert, 1987). Accordingly, introns are viewed as ancient features which have been retained in eukaryotes but which can also sometimes be lost. The 'intron late' theory states that the introns were acquired by early eukaryotic genomes and that introns are derived from transposable elements (Hickey, 1992; Cavalier-Smith, 1991a; Dibb and Newman, 1989). According to this model, early genes were continuous coding regions which were invaded by introns (transposable elements) at sites having a consensus sequence with the borders of introns. It is important to note that both models allow for exon shuffling later in eukaryote evolution.

According to the 'introns early' scenario, the ancestral actin gene would have to contain 40 introns and the evolution of the actin multigene family within and between

individual eukaryotic lineages would be marked by selective loss of introns. Dibb and Newman (1989) have argued against such a progenitor actin gene containing that many introns. The relationship between intron positions in actin genes and protein structure has been studied by Weber and Kabsch (1994). The pattern of intron locations were found not to be correlated with any structural or functional properties of actin. In fact, the positions of introns in actin sequences appear to be uniformly distributed along the amino acid sequences. These results are inconsistent with the predictions of the 'introns early' theory. They do support the idea that introns, at least in actin, were acquired late and that their acquisition was unrelated to the protein structure (Weber and Kabsch, 1994). Similar results have been obtained for other proteins (Stoltzfus et al., 1994).

Intron loss in angiosperm actin genes

In this study three plant actin clones were found which did not possess the three introns typical of plant actin genes. One of these genes, the tomato clone Tm32, lacked the second intron, and the two potato clones, Pt42 and Pt65, lacked both the second and third introns. The loss of the introns was precise such that the reading frame was not altered. Since no mechanism is known by which nuclear pre-mRNA introns can be spliced out at the DNA level (Perlman et al., 1990), it is assumed that the introns were removed at the RNA level by the usual means, reverse transcribed into a cDNA copy which then was used in a gene conversion event with an actin gene in the chromosomes. If the conversion event spanned the location of the intron, the result would be the precise removal of the intron from the chromosomal copy.

The involvement of cDNAs in homologous gene conversion events resulting in the precise splicing out of an intron has been clearly demonstrated in yeast (Derr and Strathern, 1993; Melamed et al., 1992), and the occurrence in animals of processed pseudogenes (Vanin, 1985), non-functional genes which lack promoter sequences and

introns and contain poly(A) stretches at their 3'end, indicate that reverse transcription of mRNAs is not uncommon. The presence of a processed actin pseudogene in potato indicates that this process also occurs in plants (Drouin and Dover, 1987). It has been noted that if gene conversion does occur through a cDNA intermediate, then genes that are highly expressed in the germ line will tend to be the main donors of genetic information (Melamed et al., 1992).

Phylogenetic analysis of these genes lacking some of the introns indicate that they may be orthologous members. If this is the case, then it is possible that the loss of the second intron in Tm32 and Pt42 and Pt65 may have had a common origin. Figure 20 shows a schematic representation of two models of intron loss in these genes. Both models are equally parsimonious since each requires only two gene conversion events leading to intron loss. If intron loss was due to homologous gene conversion events, then model 1 assumes that the loss of the second intron did not result in a nonfunctional gene, since its product would be needed for a second gene conversion event in the potato genes. Since gene conversion has been observed between genes which are up to 15% different at the nucleotide level, this assumption is not needed (Harris et al., 1993). Therefore, any other member of the actin family being expressed in tissue leading to the germ line could potentially be a donor.

In order for mutations to be fixed in a population or a species, they must first be passed from generation to generation. In animals, these mutations must occur in the germline, and not the somatic tissue. In plants, the germline is derived from somatic tissue late in plant development. Therefore, mutations will be fixed if they occur in germ tissue or tissue leading up to it. Phylogenetic analysis in this study showed that the tomato and potato genes lacking some of the introns always cluster with the tobacco gene, Tob1, suggesting that they may be orthologous. It is interesting therefore, that Tob1 has been shown to be highly expressed in pollen (Thangavelu et al., 1993).

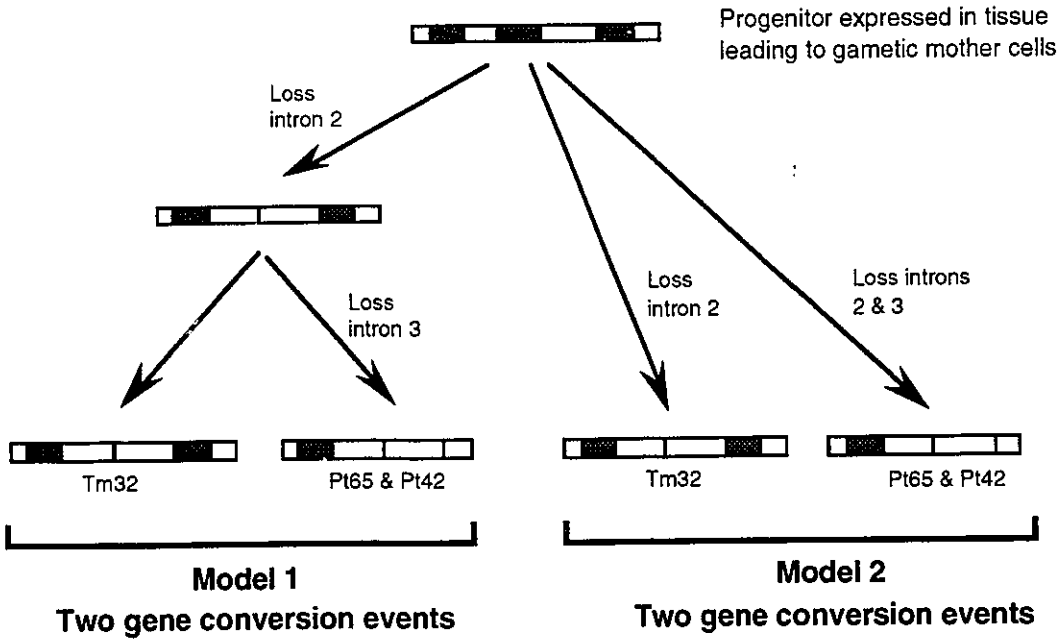

Figure 20: Schematic representation of two models of intron loss for putative orthologous genes in potato and tomato. Loss of introns 2 and 3 in model two represent a single gene conversion event. Shaded boxes represent introns. Open boxes represent exons. Introns and exons are not drawn to scale.

Assuming that there is a conservation in function between orthologous genes, these results support the notion that the progenitor gene was also expressed in the germ tissue. Whether or not Tm32, Pt42 and Pt65 are functional and expressed in the germ tissue is unknown. The tomato gene, Tm32, has frameshift mutations and an intron which is shorter than that expected to be efficiently spliced out, so it is presumed to be a pseudogene. The potato genes, Pt65 and Pt42, contain an in frame insertion and deletion, respectively, and could be functional. Work is now underway to acquire potato pollen RNA for northern analysis to determine if these genes are expressed. If they are expressed, it will have some bearing as to why intron positions are so highly conserved in angiosperm actin genes.

Although the mechanism of intron loss, as discussed above, is presumed to have taken place through a gene conversion event, an analysis of dicot actin genes failed to detect any gene conversion events, especially involving Tm32, Pt65 and Pt42. This could be because the progenitor gene was not cloned and therefore not available for analysis. Alternatively, intron loss could have been the result of homologous gene conversion events, which cannot be detected since the progenitor gene itself would be involved.

GC Content and Codon usage

The genes and genomes of prokaryotes and eukaryotes have been found to vary widely in their G+C content (Yoshida and Hickey, 1995; Lloyd and Sharp, 1992; Andersson and Kurland, 1991; Campbell and Gowri, 1990; Murray et al., 1989; Aota and Ikemura, 1986; Sharp and Li, 1986; Bennetzen and Hall, 1982). In angiosperms, the G+C content varies from as much as 80% in monocots to as low as 30 % in dicots (Meagher et al., 1989). Our results show that the G+C content of actin genes in dicots is significantly lower than those of monocots. Two previous studies have also showed

that there is a significant difference in G+C content between dicot and monocots genes (Campbell and Gowri, 1990; Murray et al., 1989).

Differences in the G+C content between genes also reflects differences in the pattern of synonymous codon usage between genes (Yoshida and Hickey, 1995; Lloyd and Sharp, 1992; Lockhart et al., 1992a, 1992b; Andersson and Kurland, 1991; Campbell and Gowri, 1990; Murray et al., 1989; Aota and Ikemura, 1986; Sharp and Li, 1986; Bennetzen and Hall, 1982). This study has shown that there is a difference in the codon usage pattern between dicot and monocot actin genes, such that the third position of codons in dicot actin genes more frequently end in A or T whereas those in monocots more frequently end in G or C. Furthermore, the results show that differences in the G+C content of the genes are only correlated with differences in G+C content of the third position of codons. This indicates that the evolution of the actin amino acid sequences are not influenced by G+C content of their respective sequences. Two previous studies have looked at the codon usage for angiosperm nuclear genes (Campbell and Gowri, 1990; Murray et al., 1989). Both studies found a clear distinction between the codon usage patterns of dicot and monocot genes. For dicots and monocots the mean average G+C at the third base of codons is 45% and 73% respectively (Campbell and Gowri, 1990). Our results agree with these observations with the exception that the G+C content at the third position of codons for monocot actins is lower than the average of a range of monocot genes. The observation that each plant species has a unique but similar codon usage bias with plants of the same taxonomic group is in agreement with Grantham's 'genome hypothesis' which states that all genes in a genome tend to have the same coding strategy in that they show a similar choice between synonymous codons (Grantham, 1980).

Interestingly, the study of Campbell and Gowri (1990) also showed that nuclear genes which are highly expressed in both dicots and monocots have a more biased

codon usage than that of genes having lower expression levels. Similarly, H3 histones which are more highly expressed appear to have a more biased codon usage (Wells et al., 1986). This situation is similar to that found in some bacteria (Andersson and Kurland, 1990; Sharp and Li 1986) as well as yeast (Bennetzen and Hall, 1982) and *Aspergillus nidulans* (Lloyd and Sharp, 1991). In these organisms the codon bias is more extreme than in plants and the codon usage patterns have been balanced with the concentrations of isoaccepting tRNA to optimize the synthesis of abundant proteins (Anderson and Kurland, 1990; Bennetzen and Hall, 1982). It has been argued that the pattern of codon usage in the highly expressed genes of these microorganisms is determined largely by selection mediated via translation whereas that of lowly expressed genes is due primarily to a relaxation of selection such that mutation pressure and random drift become more important (Andersson and Kurland, 1990; Sharp and Li, 1986). More comparisons of the pattern of synonymous codon usage between highly expressed and lowly expressed plant genes need to be done before more firm conclusions can be drawn about any similarity in codon usage with microorganisms. In the case of plant actin genes, there is not enough information on the relative levels of expression of individual genes to make any correlation with codon usage patterns.

The genome of mammals seems to be organized into patches of DNA having varying G+C contents. The G+C content of mammalian genes, and thus their synonymous codon usage patterns, appears to be determined by their location in the genome such that the G+C content of a gene corresponds to that of the surrounding DNA (Aota and Ikemura, 1986). Thus the selective constraints on the third position of codons appear to be relaxed in mammalian genes and probably the most dominant factor influencing the G+C content in these genes is a mutational bias which differs along different tracts of DNA. It has been reported that the G+C content at the third position of codons in monocot genes most closely resembles that of mammalian genes

(Campbell and Gowri, 1990). However, our study shows that, unlike the situation in animals, there is only a weak correlation between the G+C content of the genes and their respective introns. This is consistent with the limited evidence suggesting that the genome of maize is not made up of patches of DNA differing in base composition (Campbell and Gowri, 1990). Thus it appears that the mechanism affecting the nucleotide composition at the third site of codons in plant genes is different from that in animals.

The pattern of codon usage is thought to be a result of selective constraints on the one hand and mutational pressure on the other (Andersson and Kurland, 1990; Sharp and Li, 1986). Thus it is likely that the forces determining the pattern of codon usage in nonfunctional genes will be predominantly caused by mutational pressure. By analysing the codon usage pattern of nonfunctional genes it might be possible to determine more accurately the nature of the nucleotide bias due to mutational pressure. Our data show that the actin gene with the greatest overall G+C content happens to be a pseudogene from maize. Since the average G+C content of the monocot actin genes appears to be lower than that for all other monocots genes, the high G+C content of the Mz65 actin gene could reflect a natural drift in the nucleotide composition as a result of biased mutation. This suggests that the mutation bias in monocots favors G and C. Since no correlation was found between the G+C content of this gene with the G+C content of the first and second position of codons, this could mean that this gene became nonfunctional relatively recently. In contrast the G+C content of the potato processed pseudogene is identical to that of the average for all dicot genes. This suggests that the mutational bias in dicots is in favor of A and T.

In some cases the nucleotide bias not only has an effect on the pattern of synonymous codon usage but also extends to nonsynonymous codon usage (Collins and Jukes, 1993). A significant correlation has been observed between changes in the

amino acid composition and the nucleotide composition at the third position of codons for alpha-amylase genes (Yoshida and Hickey, 1995). From a phylogenetic perspective, this means that in such cases, phylogenetic trees based on amino acid sequences will not be devoid of the effects of nucleotide bias. Our studies suggest that the bias in the G+C content between the members of the plant actin gene family does not affect their amino acid sequences. This is expected since the actin amino acid sequences in all eukaryotes are highly conserved and consequently selective constraints operating at nonsynonymous sites against nucleotide mutation pressure will be high. By removing the third position of codons from our sequence alignments, it was possible to remove any nucleotide bias which may effect the phylogenetic trees. It has been shown that differences in the G+C content at the third position of codons result in variation among genes in the rate of synonymous substitution, such that synonymous substitution rate is inversely proportional to the degree of codon usage bias (Sharp and Li, 1986). In a study of H3 histone genes, Wells et al. (1986) found that members of this gene family clustered on a phylogenetic tree because they shared a similar synonymous codon bias. The effect of G+C content on phylogenies has also been extensively studied by Lockhart et al. (1992a,b)

Variation in actin amino acid sequence

This study has shown that members of the plant actin multigene families may differ from one another by as much as 14% in their amino acid sequence. To put this variation into perspective, the percent amino acid differences were determined for actins of other groups of eukaryotes (Figure 21). Put into a broader context, the variation observed within plants exceeds only that seen in animals, suggesting that either the actin multigene family in plants began to diverge earlier than that of animals, or that they are evolving faster. The variation seen in plants is similar to that observed for

fungal actins. Based on the calculated rates of evolution as well as the distance-based eukaryotic actin tree (Figure 16), it is concluded that the higher variation in amino acid sequence in plant and fungal actin genes relative to those in animals is due to a faster rate of evolution rather than an earlier divergence (see below).

The protists represent a heterogeneous assemblage of eukaryotes which have for the most part been classified by exclusion from other major eukaryotic lineages and whose phylogenetic relationships are not well defined (Margulis and Schwartz, 1988). Based on the phylogenetic relationships in Figure 16, protists were split into three groups, protist 1, protist 2 and slime molds, in recognition of the deep divergence existing within them.

By far the greatest variation in actin amino acid sequence is observed within group 1 protists. Most of this divergence being attributable to the inclusion of *Giardia lamblia* and *Euploites crassus* in this group (see Figure 16). Such a high degree of amino acid sequence divergence confirms the idea that most of the variation among eukaryotes exists within the protists and that divergences within this group are very ancient (Cavalier-Smith, 1991b). Such divergent actins might also have functions which are different from those of other eukaryotes which have evolved more recently (i.e. animals, plants and fungi).

The divergence in actin amino acid sequence between different eukaryotic lineages indicates that the actins of fungi, slime molds and group 2 protists are more similar to those of animal actins than those of plants. The close similarity of group 2 protists to animals is due to the inclusion of *Entamoeba* and *Acanthamoeba* actins in this group, both of which cluster with animals to the exclusion of plants in the eukaryotic actin tree shown in Figure 16. The greatest intergroup variation is observed between group 1 protists and the other lineages. In addition, the variation between group 1 protist and the other groups is roughly the same suggesting an equal time since

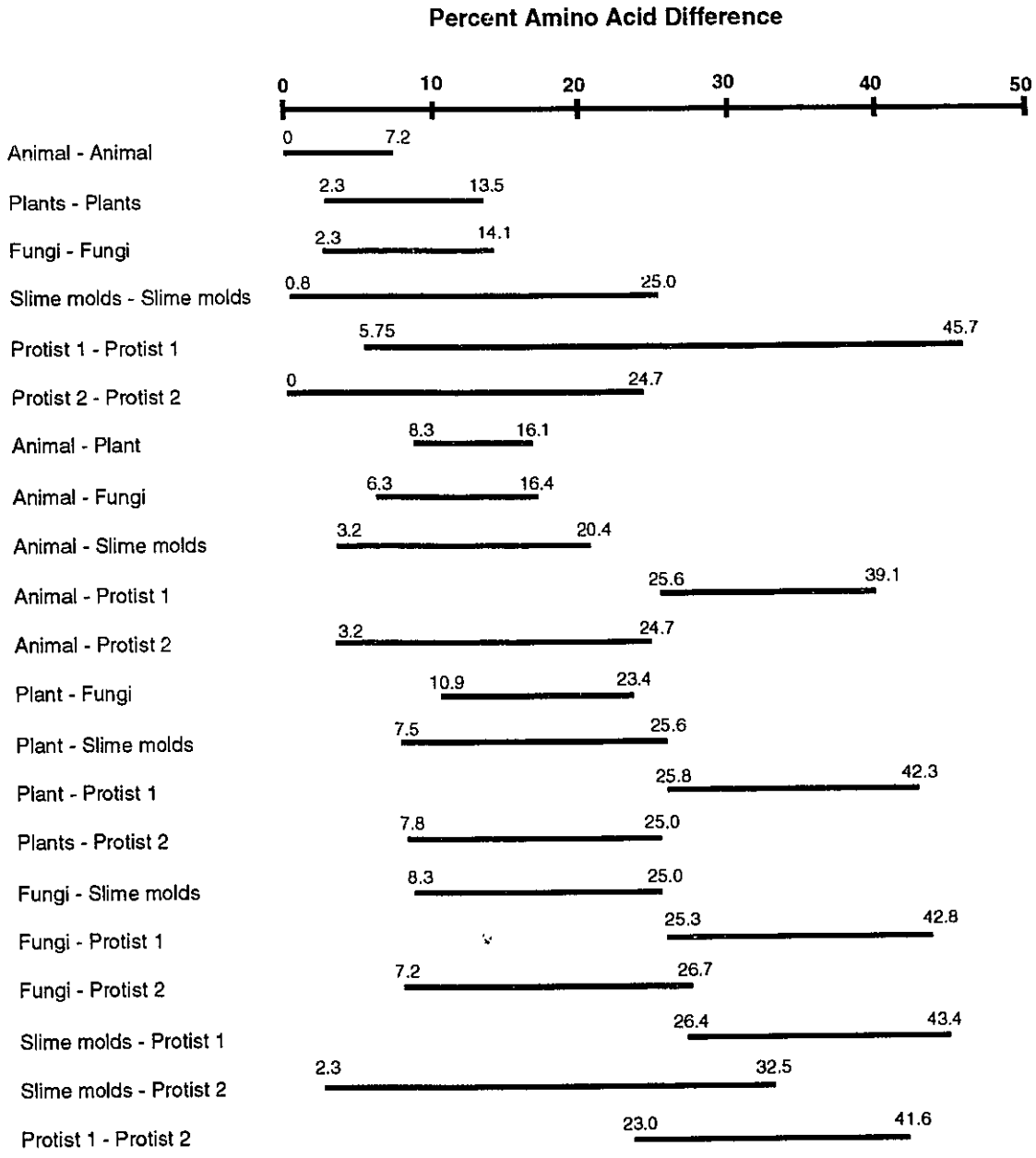

Figure 21: Bar graph showing the percentage of observed amino acid differences between actin genes in various taxa. Sequences included were those shown in Tables 1 and 3. The *Volvox* actin gene was not included.

the divergence of group 1 protists and the lineage(s) leading to the other eukaryotic groups (see below).

Phylogeny of angiosperm actin genes

This work has contributed significantly to the number of actin genes available for use in the phylogenetic study of the actin multigene family in plants. A total of 41 new actin genes were isolated from potato, tomato, tobacco, soybean, and maize. Of these 41 unique actin clones, 32 were used, in addition to other plant actins already available from the Embl/Genbank databases, for phylogenetic analysis. The other nine clones were either not completely sequenced or turned out to be truncated clones not usable for phylogenetic analysis since they would limit the size of the data set.

Phylogenetic analysis in this study, using both parsimony and distance methods, have shown that the evolution of the actin multigene family in plants is quite complex, marked by many ancient and recent gene duplication events. Many of the branches include genes from both soybean as well as those of Solanaceae species. Therefore, gene duplications leading to these various branches must have occurred prior to the divergence of the lineages leading to the *Solanaceae* and soybean, more than 110 million years ago. Using 750 MY as the time since the last common ancestor shared by land plants and *Volvox*, and taking into consideration the branch lengths from the distance-based DNA trees, the earliest gene duplications in the plant actin multigene family are estimated to have taken place starting at about 300-400 MY ago. Since the age of angiosperms is thought to be more recent (Wolfe et al., 1989; Brandl et al. 1992), this implies that many of the gene duplication events in the actin multigene family took place prior to the origin of the angiosperms, and implies that orthologous genes will be identified in non-angiosperms as well. A preliminary analysis of actin genes from two gymnosperms (Cycad and *Podocarpus*) and a fern (*Osmunda*) suggest

that this might be the case (Moniz de Sá and Drouin, unpublished). In addition, a partial sequence of an actin gene from pine clusters with a soybean actin gene to the exclusion of other angiosperm actins (Reece et al., 1992; Kenny et al., 1988). On the other hand, there is some indication that the origin and duplication of the actin multigene family in ferns may be independent of that in monocots and dicots (Bhattacharya, personal communication). However, the possibility that concerted evolution in fern actin genes is masking orthologous origins has not been ruled out. Also, since polyploidy is very common in ferns (some ferns have a diploid chromosome number exceeding 800) (Soltis et al., 1992), it is possible that the apparent monophyly of fern actin genes may be the result of actin gene loss.

In this study, 38 dicot and 13 monocot actin genes were used to construct phylogenetic trees which group plant actin genes into roughly three subfamilies (although weakly supported by bootstrap values). Previous phylogenetic studies, using a much more limited number of sequences and with monocots being poorly represented, also grouped plant actin genes into three diverse subfamilies (Meagher, 1991; Drouin and Dover, 1990). While the composition of each subfamily differs a little from those of previous studies, this difference is mainly attributable to the fact that the number of sequences used in this study is six times higher than previous analysis of plant actin genes (Reece et al., 1992; Meagher, 1991; Drouin and Dover, 1990). The inclusion of more monocot actin sequences has shown that most of them cluster on the tree to the exclusion of dicot genes. Since, as mentioned above, the divergence of the actin multigene family predates that of the monocot-dicot split, it is likely that concerted evolution has played a role in the evolution of monocot genes (see below).

The genes cloned in this study as well as those previously known are all well represented in all the major branches of the phylogenetic trees. This suggests that the published actins as a whole may already be representative of most of the variation

existing within the plant actin multigene family. The addition of 32 new actin genes has simply added more detail, i.e. more short branches to the plant actin tree. However, given that these new actin genes were PCR amplified using degenerate primers designed on the basis of plant actin genes already in the data base, it could be that the fact that no new major branches were identified on the phylogenetic trees is due to the possible bias introduced by these PCR primers.

Increasing the number of actin genes for phylogenetic analysis has resulted in the identification of putative orthologous genes among the four dicot species. This is the first time that orthologous actin genes have been identified among plants. The previous suggestion that the *Arabidopsis* gene, Aac1, and the rice gene, Rac1, may be orthologous is not supported by any of the phylogenetic trees presented here (Reece et al., 1992). The identification of orthologous genes is important because they can be used to determine the rate of gene evolution if the species divergence times are known (see below). In addition, orthologous genes can be used to test hypotheses concerning the conservation of function. For a large diverse multigene family like actin it would be interesting to determine if orthologous members in divergent species have maintained the same function or if there are regulatory shifts, as reported for glutamine synthase genes in angiosperms (Doyle, 1991).

Actin genes not appropriate to study plant phylogeny

One conclusion which can be drawn from this study is that actin genes are not suitable for use in elucidating the phylogenetic relationships of land plants. There are four main reasons for this.

First, actin genes are too highly conserved, and given the size of the coding sequence, 377 amino acids, as well as the time scale involved in the evolution of land plants (less than 450 MY), not enough variation has accumulated in these sequences to

give unambiguous phylogenetic results. This is clearly illustrated in all the phylogenetic trees, regardless of method, where there was very little support for most of the deep branches, as deduced from their low bootstrap values.

Second, the actin multigene family underwent many gene duplication events early in the evolution of land plants. The highly conserved nature of actins compounds this problem resulting in many deep nodes which are separated by very small distances. As a consequence, there will be a lot of stochastic errors resulting in phylogenetic trees with very low bootstrap values. In some cases the distances separating the deep nodes are smaller than the standard error.

Based on molecular evolutionary studies, it now appears that angiosperms had their origins much earlier than previously thought, perhaps as early as 200-350 MY ago (Brandl et al., 1992; Wolfe et al, 1989; Martin et al., 1989). The angiosperms may have occupied upland habitats which were not conducive to fossilization and, consequently, were not detected in the fossil record until much later when they expanded their range (Cleal, 1989). In addition, seed plants may have shared a common ancestor as late as 300 MY ago, a date substantially more recent than previously thought (Savard et al., 1994). If this is the case, then the time frame separating the origins of the various seed plants will be shorter, meaning that using a highly conserved protein such as actin to infer the phylogenetic relationships among many seed plants will not provide the resolution needed. Thus, the patterns of gene duplication as well as the pattern of speciation are important factors in choosing a gene for phylogenetic study. The early radiation of the plant actin multigene family as well as its conserved nature suggest that it will not be useful to infer unambiguously the phylogenetic relationships of land plants.

Third, the plant actin multigene family is too large to be practical for use in inferring phylogenetic relationships among land plants. Inferring species relationships

from gene trees assumes that orthologous genes are being compared. The unambiguous identification of actin orthologous genes from a variety of plant species, especially given the problems mentioned above, will probably be problematic, costly and time consuming. Thus, another protein, encoded by a single copy gene or by a much smaller gene family, whose orthology among various species is unambiguous and whose rate of evolution is appropriately faster, will be of better use for studying the phylogenetic relationships of land plants.

Finally, we now know that there is gene conversion between multigene family members. This will further confound phylogenetic analysis by obscuring orthologous relationships (see below).

Actin as a phylogenetic marker for studying eukaryotic evolution

The use of actin amino acid sequences to infer the phylogeny of eukaryotes has reaffirmed the status of plants, animals and fungi as monophyletic groups. In addition, the actin tree supports the notion that fungi and animals are more closely related to each other than either is to plants. A recent phylogenetic analysis based on four different nuclear genes (actin, α-tubulin, β-tubulin, and elongation factor 1α) (Bhattacharya and Ehlting, 1995; Hashimoto et al., 1994; Baldauf and Palmer,1993; Hasegawa et al., 1993), as well as 18S rRNA (Cavalier-Smith et al., 1994; Wainright et al., 1993; Bhattacharya et al., 1991; Sogin, 1989) have resulted in a similar conclusion. The status of animals and fungi as sister groups is supported by other shared characters, such as motile cells with basal flagella, similar mitochondria with reduced gene content and similar codon assignments, lack of chloroplasts, chitinous cytoskeleton, and the use of glycogen for carbohydrate storage (Baldauf and Palmer, 1993).

The phylogenetic relationships among plants, fungi, and animals remains controversial. Sidow and Thomas (1994) have carried out a phylogenetic analysis on the two largest subunits of RNA polymerase II and concluded that animals group with plants to the exclusion of fungi. However, their study is flawed since both fungi and plants were severely under-represented and no mention is made of the possible error introduced into their analysis due to a GC content bias. Another study supporting plant-animal monophyly was based on multiple lines of molecular evidence used rRNA, tRNA, and six protein sequences (Gouy and Li, 1989). This study has been criticized by Baldauf and Palmer (1993) as being limited by the narrow representation of major taxonomic groups.

Plants, animals, fungi, and some protists, such as slime molds, represent relatively recent evolutionary lineages with respect to the protists. Plants, animals, fungi and slime molds appear to have evolved almost simultaneously. The relatively little distance associated with their origins in part explains why there has been discrepancies in the relationship among animals, plants and fungi (Baldauf and Palmer,1993; Gouy and Li, 1989). In many trees, plants come out as the sister group of animals to the exclusion of the fungi. These likely erroneous results are probably due to GC content bias or due to having very few representatives of each of these major lineages.

The phylogenetic relationship of the mitochondria-lacking gut parasite, *Entamoeba histolyca*, remains uncertain. A phylogenetic tree based on Elongation Factor 1α could not firmly establish its relationship to other protists (Hashimoto et al., 1994). The actin-based tree from this study is in accordance with the rRNA phylogenies (Sogin 1989) in suggesting that *Entamoeba histolyca* diverged after the kinetoplastid (*Leishmania major*, and both Trypanosome species) and ciliated protozoan

(Tetrahymena, Euploites, and Oxytricha) species. This supports the idea that Entamoeba has secondarily lost its mitochondria (Schlegel, 1994) and fully justifies Cavalier-Smith's removal of the Entamoedidae from the Archezoa, all of which never gained a mitochondrion through a symbiosis event (Cavalier-Smith, 1991).

The phylogenetic status of *Dictyostelium discoideum*, a cellular slime mold, has also been controversial. Phylogenies based on rRNA suggest that this species diverged before the lineage leading to ciliates, plants, fungi and animals (Sogin, 1991; Hendricks et al., 1989; Gunderson et al., 1987). In fact, the sequence of the small-subunit rRNA molecule of *D. discoideum* has often been used as an outgroup to root the rRNA phylogenies of ciliates, plant, fungi, and animal phylogenies (Wainright et al., 1993; Field et al., 1988). In contrast, phylogenies based on protein sequences suggest that *D. discoideum* is not an outgroup to the lineage leading to plants, fungi and animals. Loomis and Smith (1990) have studied phylogenies based on eight protein sequences and concluded that *D. discoideum* is more closely related to animals and plants than to fungi, and that this species diverged from the line leading to animals at about the same time as the plant/animal divergence. The discrepancy between rRNA and protein based trees has been attributed to an erroneous placing in the rRNA trees due to the relatively high A+T content of the small-subunit rRNA genes of this species (Loomis and Smith, 1990). A phylogenetic study of the elongation factor 1α also suggests an origin of the lineage leading to *D. discoideum* comparable to that of plants, fungi and animals (Hasegawa et al., 1993). The actin amino acid tree from this study indicates that the slime molds (*D. discoideum* and *P. polycephalum*) diverged after the lineage leading to plants, but before the lineage leading to fungi and animals. Similar results have been previously reported based on an analysis of amino acid sequences as well as of first and second codon positions of actin genes (Baldauf and Palmer, 1994; Bhattacharya et al.,

1991), although a more recent analysis suggested that a slow-down in the rate of evolution *of D. discoideum* genes might be responsible placing it at the crown of eukaryote evolution (Bhattacharya and Ehlting, 1995). The recent report of *Dictyostelium discoideum* having its NAD-reducing hydrogenase gene in its mitochondria whereas plants, animals, and fungi have this gene encoded in the nucleus provides strong support for D. *discoideum* having diverged before the lineage leading to plants, animals, and fungi (Cole et al., 1995).

It is likely that actin sequences will be very useful for elucidating the phylogenetic relationships between the major eukaryotic lineages. This is reflected in the relatively high bootstrap values associated with the deep nodes in the eukaryotic actin tree, in contrast to the angiosperm actin trees. The higher bootstrap values are due to a greater actin sequence variation within eukaryotes, making actins phylogenetically more informative at the level of eukaryotes than at the angiosperm level. In addition, it is likely that the actin gene within each of the major eukaryotic lineages are orthologous since they underwent duplications after the major eukaryotic lines were already established. Thus, although actin genes exist as multigene families in most eukaryotes, actin gene duplications are likely to be "irrelevant to the branching order among major groups" (Baldauf and Palmer, 1994).

Gene conversion in maize actin genes

Gene conversion is the non-reciprocal exchange of genetic information between closely related DNA sequences and is a well documented phenomena in eukaryotes (Arnheim, 1983). The net effect of gene conversion is the sequence homogenization of gene family members. Gene family members within a taxa which are undergoing gene conversion will, depending on the extent of sequence homogenization, cluster together on a phylogenetic tree rather than with their orthologues from other taxa (Meagher et

al., 1989). Thus, gene conversion can obscure true evolutionary relationships. Gene conversion has been implicated between two actin genes in sea urchin (Durica et al., 1988; Crain et al., 1987). A previous study on potato actin genes failed to detect any gene conversion (Drouin and Dover, 1990).

The phylogenetic analysis of plant actin genes presented here showed that, despite the fact that the actin multigene family underwent several gene duplication events which preceded the monocot-dicot divergence, most of the monocot actin genes clustered together on the tree and no obvious orthologues to dicot actin genes were found. In contrast, the actins from the various dicot species were well represented in the other major branches of the phylogenetic tree. The clustering of the monocot genes, together with the relatively long branches separating most of the monocot genes from one another, suggests that these sequences may have undergone concerted evolution early in the evolution of monocots. Two gene conversion events were detected between the maize actin genes, one between Mz56 and Mz63, and another between Mz56 and Mz81. Both gene conversion events occurred relatively recently. However, the algorithm used to identify gene conversion events is supposedly very stringent (Frederic Pratt, personal communication) and it is possible that other gene conversions have occurred but were not detected.

Rates of actin gene evolution

In this study, the relative rate test was used to show that the rate of non-synonymous substitution of actin genes is constant in different angiosperm taxa. Rate differences at synonymous sites could not be determined using *Volvox* as outgroup due to saturation at these sites. Rate differences have been widely reported, especially between different animal groups (reviewed in Li, 1993b), as well as between mitochondrial, chloroplast and nuclear genomes (Wolfe et al., 1987). In plants, studies

of the chloroplast encoded gene, ribulose-1,5-biphosphate carboxylase large subunit (*rbc*L) (Bousquet et al, 1992; Frascaria et al., 1993), as well as other chloroplast genes (Wolfe et al., 1987) have shown that considerable heterogeneity exists in the rates of synonymous and non-synonymous substitution in plants. In general, it was found that the rate is faster in monocots than in dicots, and higher in annuals than in perennials. In addition, the rate heterogeneity is more pronounced at non-synonymous than synonymous sites. However, lineage specific rate differences have not been detected among plant nuclear encoded genes (Savard et al., 1994; Bousquet et al., 1992; Wolfe et al., 1987).

The rates of synonymous and non-synonymous substitutions were calculated from 18 independent comparisons of sequence divergence between orthologous genes separated in time for 40 and 110 MY. Although the putative plant actin orthologues were identified on the basis of their position on the phylogenetic trees, several other pieces of information are consistent with their status as orthologues. These include shared intron loss, intron sequence similarity, as well as the similar values obtained from the independent estimates of the rates of evolution. However, these hypotheses concerning orthology in plant actins need to be tested in order to give strength to conclusions derived from their use. Work is currently underway using southern blot analysis to confirm these hypotheses.

The rate calculations indicate that while the average synonymous rate of substitutions (5.44×10^{-9}/site/year) is comparable to that observed for animal (4.61×10^{-9}/site/year) (Li and Tanimura, 1987; Li and Graur, 1991) as well as plant ($5.0\text{-}30.0 \times 10^{-9}$/site/year)(Wolfe et al., 1987) nuclear genes. The average non-synonymous rate of substitution (0.13×10^{-9}/site/year) is 4-10 times faster in dicot than in animal actin genes ($0.01\text{-}0.03 \times 10^{-9}$/site/year) (Li and Graur, 1991). The results presented here show that the rate of non-synonymous substitution is the same for both dicot and fungal actin

genes. The faster rate in plant versus animal actin genes is also apparent from the eukaryotic actin distance tree (Figure 16). These results are contradictory to the previously held belief that actin genes are evolving at the same rate in both plants and animals, and that the greater variation in amino acid sequence observed among plant actins is due to the more ancient origins of the actin gene family in plants than in animals (Reece et al, 1992; Hightower and Meagher, 1986). Hightower and Meagher (1986) estimated the rate of evolution using the divergence of 1000 MY for all kingdoms and thus obscured rate differences which have taken place more recently in the evolution of plant and animals because of an averaging effect. A similar observation was encountered when conducting the relative rate test between actins of various "higher" eukaryotic groups using *Fucus disticus* as an outgroup (Table 12).

The greater variation in amino acid sequence within the plant actin multigene family is due the faster rates of evolution rather than a more ancient origin with respect to that in animals. The fact that the actin multigene family in plants may have begun to diversify as early as 300-400 MY ago is not consistent with previous explanations for the diversity of actins in plants. It has previously been proposed that the diverse actin gene types present today arose concomitantly with the numerous cell types present in vascular plants, and that their evolution has been preserved throughout vascular plant evolution because they have unique patterns of actin gene regulation or encode proteins with unique functions (Meagher, 1991; Hightower and Meagher, 1986). However, the results from this study indicate that the diversity of actin types observed today occurred well after the divergence of vascular plants. In fact, the higher rate of evolution of plant actin genes suggest nonconservation rather than conservation in function, as observed for cytoplasmic and muscle actins in vertebrates. It is not clear why plants would need more diverse actins performing more diverse functions as opposed to animals which have more varied tissue types and yet possess fewer functional actins with lower

sequence diversity (Reece et al, 1992; Shetterline and Sparrow, 1994). While it is true that different subfamilies of actin genes are differentially regulated (Wolf and Timko, 1994; Thangavelu et al., 1993; Meagher, 1991; McLean et al., 1990), there might not be a strong conservation in function spanning the whole range of plant actin gene evolution, as previously suggested. Preliminary results suggest that the fern actins are all constitutively expressed in all tissues (Bhattacharya, personal communication). More actins need to be isolated from nonvascular and early vascular plants and their expression patterns studied to address this issue more fully. However, it is likely that the diversity of actin functions is mediated by interactions with the diverse actin-binding proteins which are known (Shetterline and Sparrow, 1994).

Using the estimated rate of non-synonymous substitutions and the estimated number of non-synonymous substitutions between putative orthologous genes (Tm41, PoAc97 and Tm52, Pt79) from potato and tomato, the divergence time of these two species has been estimated to be 21 ± 12 MY. A close relationship between potato and tomato has previously been documented based on morphological and molecular data. (Borisjuk et al., 1993; Martin et al., 1986; D'Arcy, 1979). In fact, it has been suggested that the tomato group should be included in the genus *Solanum* rather than as separate genus *Lycopersicon* (Borisjuk et al., 1993 and references therein).

Summary

Actin is a contractile protein and a major component of the eukaryotic cytoskeleton, and is found only in eukaryotes. In most species, actin is encoded by a multigene family. The evolution of this gene family has been well studied in animals, where much is also known about the expression patterns and the function of the various actin isotypes encoded by the different gene family members. In contrast, our knowledge of the actin multigene family in plants is relatively limited. Although it is known that the actin multigene family in plants is generally larger and their amino acid sequences are more diverse than that in animals, few representatives have been isolated from any one species. In addition, very little is known about the patterns of expression and functions of the various members of the plant actin multigene family.

This study has focused on the evolution of the actin multigene family in a group of taxonomically well-defined angiosperms in hopes of understanding the patterns and processes involved in their evolution. Using the polymerase chain reaction, a total of 44 unique actin genes have been cloned and sequenced from potato, tomato, tobacco, soybean, and maize. Phylogenetic analysis of 32 of these genes along with those previously published actin genes indicate that angiosperm actin genes are monophyletic and that they can be divided roughly into three subfamilies which diverged from each other prior to the divergence of monocots and dicots. In addition, the evolution of the actin multigene family has been marked by many gene duplication events resulting in a high copy number of actin genes within each species (20-40 copies).

The evolution of the actin multigene family in monocots has been shaped in part by gene conversion events resulting in concerted evolution. Two relatively recent gene conversion events were detected between maize actin genes. However, the fact that one of the major subfamilies of actins identified here (which diverged prior to the monocot-

dicot divergence) was composed exclusively of monocot actin genes suggests that earlier gene conversion events occurred which can no longer be detected and which are responsible for forcing the monocot and dicot sequences apart on the phylogenetic trees. Thus, the concerted evolution of actin genes in monocots has precluded the identification of orthologous genes between monocots and dicots, and consequently no inference on the time of divergence between monocots and dicots could be made.

The phylogenetic analysis presented here has identified six sets of putative orthologues between dicot species whose approximate times of divergence are known. These putative orthologues were used to calculate the rates of evolution. The synonymous rate (5.44×10^{-9}/site/year) is similar to that of other nuclear protein-encoding genes, but the non-synonymous rate (0.13×10^{-9}/site/year) is 4-10 times higher than that observed in vertebrate actin genes. This suggests that the greater amino acid sequence diversity observed between plant actin genes relative to those of animal actin genes is due, not to their more ancient origins, as previously proposed (Meagher, 1991), but their faster rate of evolution. Thus the diversification of actin isotypes in plants was not concomitant with, nor a prerequisite for, the diversification of the various tissue types during land plant evolution.

This study has shown that actin genes will be of limited use in determining the phylogenetic relationships of land plants. This is because (1)the actin gene family is very large, making identification of orthologues difficult, (2) members of this family are undergoing concerted evolution, and (3) actin genes are too conserved and do not give the resolution needed at this level of study. In contrast, actin genes will likely be of use for inferring the phylogenetic relationships between different eukaryotic lineages, especially protist lineages which diverged early during the evolution of eukaryotes.

The work presented here on the evolution of angiosperm actin genes has laid down the ground work for further characterization of the evolution and function of this

large and diverse gene family. The identification of six sets of actin orthologues from angiosperm species of varying divergence times should enable one to address many unanswered questions. For example, why are plant actin genes under less evolutionary constraints, as deduced from their faster rates of evolution, than those of vertebrates? Does the greater diversity of plant actin amino acid sequences, relative to those of animals, mean that actins perform more diverse functions in plants than animals? Are the members of this gene family differentially expressed and if so, are these patterns of expression conserved between orthologues? Most intriguing of all, the functions of actin are so many, so complex, and so fundamental to eukaryotic life, and yet in some organisms (many fungi for example) actin is encoded by only one gene and in others encoded by large multigene families. Obviously there is still a lot we do not know about the evolution and function of actins.

REFERENCES

Andersson, G.E., and Kurland, C.G. (1990). Codon preferences in free-living microorganisms. *Microbiological Reviews* **54(2):** 198-210.

Andersson, G.E., and Kurland, C.G. (1991). An extreme codon preference strategy: codon reassignment. *Molecular Biology and Evolution* **8(4):** 530-544.

Aota, S., and Ikemura, T. (1986). Diversity in G+C content at the third position of codons in vertebrates genes and its cause. *Nucleic Acids Research* **14(16):** 6345-6355.

Arnheim, N. (1983). Concerted evolution of multigene families. In: M. Nei and R.K. Koehn (Eds). *Evolution of genes and proteins.* Sinauer Associates.

Arumuganathan, K., and Earle, E.D. (1991). Nuclear content of some important plant species. *Plant Molecular Biology Reporter* **9(3):** 208-218.

Baird, W.V., and Meagher, R.B. (1987). A complex gene superfamily encodes actin in petunia. *The EMBO Journal* **6(11):** 3223-3231.

Baldauf, S.L., and Palmer, J.D. (1993). Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proceedings of the National Acadamy of Sciences, USA* **90:** 11558-11562.

Beach, R.L., and Jeffery, W.R. (1992). Multiple actin genes encoding the same α-muscle isoform are expressed during ascidian development. *Developmental Biology* **151**: 55-66.

Beifuss, M.J., and Durica, D.S. (1992). Sequence analysis of the indirect flight muscle actin-encoding gene of *Drosophila simulans. Gene* **118**: 163-170.

Bennetzen, J.L., and Hall, B.D. (1982). Codon selection in yeast. *The Journal of Biological Chemistry* **257(6):** 3026-3031.

Bernatzky, R., & Tanksley, S.D. (1986a). Genetics of actin-related sequences in tomato. *Theoretical and Applied Genetics,* **72**:314-321.

Bernatzky, R., & Tanksley, S.D. (1986b). Majority of random cDNA clones correspond to single loci in the tomato genome. *Molecular and General Genetics* **203:** 8-14.

Berbee, M.L., and Taylor, J.W. (1993). Dating the evolutionary radiations of the true fungi. *Canadian Journal of Botany* **71:** 1114-1127.

Bhattacharya, D., and Stickel, S.K., (1994). Sequence analysis of duplicated actin genes in *Lagenidium giganteum* and *Pythium irregulare* (Oomycota). *Journal of Molecular Evolution* **39**: 56-61.

Bhattacharya, D., Stickel, S.K., & Sogin, M.L. (1991). Molecular phylogenetic analysis of actin genic regions from *Achlya bisexualis* (Oomycota) and *Costaria costata* (Chromophyta). *Journal of Molecular Evolution,* **33**:525-536.

Bhattacharya, D., and Ehlting, J., (1995). Actin coding regions: gene family evolution and use as a phylogenetic marker. *Archive Protistenkunde*, in press.

Borisjuk, N., Borisjuk, L., Petjuch, G., and Hemleben, V. (1993). Comparison of nuclear ribosomal RNA genes among *Solanum* species and other *Solanaceae. Genome* **37**: 271-279.

Bousquet, J., Strauss, S.H., Doerksen, A.H., and Price, R.A. (1992). Extensive variation in evolutionary rate of *rbcL* gene sequences among seed plants. *Proceedings of the National Academy of Sciences, USA* **89**:7844-7848.

Brandl, R., Mann, W., and Sprinzl, M. (1992). Estimation of the monocot-dicot age through tRNA sequences from the chloroplast. *Proceedings of the Royal Society of London B* **249:** 13-17.

Brown, B.W., Brauner, C., Chan, A., Gutierrez, D., Herson, J., Lovato, J., and Posley, J. (1993). STPLAN vs 4.0: claculations for sample sizes and related problems. University of Texas.

Campbell, W.H., and Gowri, G. (1990). Codon usage in higher plants, green algae, and cyanobacteria. *Plant Physiology* **92**:1-11.

Cavalier-Smith, T. (1993). Kingdom protozoa and its 18 phyla. *Microbiological Reviews* **57(4):** 953-994.

Cavalier-Smith, T. (1991a). Intron phylogeny: a new hypothesis. *Trends in Genetics* **7(5):** 145-148.

Cavalier-Smith, T. (1991b). The evolution of cells. In: S Osawa and T Honjo (Eds), *Evolution of life: fossils, molecules and culture.* Springer-Verlag Tokyo. pp 271-304.

Cavalier-Smith, T. (1991c). Archamoebae: the ancestral eukaryote? *Biosystems* **25:** 25-38.

Cavalier-Smith, T., Allsoppp, M.T.E.P., and Chao, E.E. (1994). Thraustochytrids are chromists, not fungi: 18s rRNA signatures of Heterokonta. *Philosophical Transactions of the Royal Society of London* B **346:** 387-397.

Cleal, C.J. (1989). Evolution in hidden forests. *Nature* **339**: 16.

Cole, R.A., Slade, M.B., and Williams, K.L. (1995). *Dictyostelium discoideum* mitochondrial DNA encodes a NADH:ubiquinone oxidoreductase subsunit which is nuclear encoded in other eukaryotes. *Journal of Molecular Evolution* **40:** 616-621.

Collins, D.W., and Jukes, T.H. (1993). Relationship between G+C in silent sites of codons and amino acid composition of human proteins. *Journal of Molecular Evolution* **36**: 201-213.

Crain, W.R., Boshar, M.F., Cooper, A.D., Durica, D.S., Nagy, A., and Steffen, D. (1987). The sequence of the sea urchin muscle actin gene suggests a gene conversion with a cytoskeletal actin gene. *Journal of Molecular Evolution* **25:** 37-45.

Cronquist, A. (1981). The evolution and classification of flowering plants. New York Botanical Garden, Bronx, NY.

Cross, G.S., Wilson, C., Erba, H.P., and Woodland, H.R. (1988). Cytoskeletal actin gene families of *Xenopus borealis* and *Xenopus laevis*. *Journal of Molecular Evolution* **27:** 17-28.

D'Arcy, W.G. (1979). The classification of the *Solanaceae*. In: The biology and Taxonomy of the *Solanaceae*. Eds J.G. Hawkes, R.N. Lester, and A.D. Skelding. Academic Press.

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1979). A model of evolutionary change in proteins. in M.O. Dayhoff, (Ed.) *Atlas of protein sequence and structure* Vol.5, National Biomedical Research Foundation, Washington, D.C.

Davidson, E.H., Thomas, T.L., Scheller, R.H., and Britten (1982). The sea urchin actin genes, and a speculation on the evolutionary significance of small gene families. In: G.A. Dover and R.B. Flavell (Eds) *Genome evolution*. Academic Press.

Derr, L.K., and Strathern, J.N. (1993). A role for reverse transcripts in gene conversion. *Nature* **361**: 170-173.

Dibb, N.J., and Newman, A.J. (1989). Evidence that introns arose at protosplice-sites. *EMBO Journal* **8:** 20150-2021.

Doyle, J.J. (1991). Evolution of higher-plant glutamine synthase genes: tissue specificity as a criterion for predicting orthology. *Molecular Biology and Evolution* **8(3):** 366-377.

Dover, G.A. (1982). Molecular drive: a cohesive mode of species evolution. *Nature* **299:** 111-117.

Dover, G.A., and Tautz, D. (1986). Conservation and divergence in multigene families: alternatives to selection and drift. *Philosophical Transaction of the Royal Society of London* B **312:** 275-289.

Drouin, G., and Dover, G.A. (1990). Independent gene evolution in the potato actin gene family demonstrated by phylogenetic procedures for resolving gene conversions and the phylogeny of angiosperm actin genes. *Journal of Molecular Evolution,* **31**:132-150.

Drouin, G., & Dover,G.A. (1987). A plant processed pseudogene. *Nature* **328**:557-558.

Durica, D.S., Garza, D., Restrepo, M.A., Hryniewicz, M.M. (1988). DNA sequence analysis and ultrastructural relationships among the cytoskeletal actin genes of the sea *urchin Strongylocentrotus purpuratus. Journal of Molecular Evolution* **28:** 72-86.

Fang, H., and Brandhorst, B.P. (1994). Evolution of actin multigene families of sea urchins. *Journal of Molecular Evolution* **39**: 347-356.

Felsenstein, J.(1988). Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* **22**:521-565.

Felsenstein, J. (1993). Phylogeny Inference Package , version 3.5 (PHYLIP). University of Washington.

Field, K.G., Olsen, G.J., Lane, D.J., Giovannoni, S.J., Ghiselin, M.T., Raff, E.C., Pace, N.R., and Raff, R.A. (1988). Molecular phylogeny of the animal kingdom. *Science* **239:** 748-753.

Flaherty. K.M., McKay, D.B., Kabsch, W., and Holmes, K.C. (1991). Similarity of the three-dimentional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein. *Proceedings of the National Academy of Sciences,* USA **88:** 5041-5045.

Frascaria, N., Maggia, L., Michaud, M., and Bousquet, J. (1993). The *rbcL* gene sequence from the chestnut indicates a slow rate of evolution in the Fagaceae. *Genome* **36:** 668-671.

Gaut, B.S., and Clegg, M.T. (1993). Molecular evolution of the *Adh1* locus in the genus *Zea. Proceedings of the National Academy of Science* USA **90:** 5095-5099.

Giannasi, D.E., Zurawski, G., Learn, G.H., & Clegg, M.T. (1992). Evolutionary relationships of the Caryophyllidae based on comparative rbcL sequences. *Systematic Botanist,* in press.

Gilbert, W. (1987). The exon theory of genes. *Cold Spring Harbor Symposia on Quantitative Biology* **52:** 901-905.

Goodall, G.J., and Filipowicz, W. (1990). The minimum functional lengh of pre-mRNA introns in monocots and dicots. *Plant Molecular Biology* **14**: 727-733.

Gouy, M., and Li, W-H. (1989). Molecular phylogeny of the kingdoms animalia, plantae, and fungi. *Molecular Biology and Evolution* **6(2):** 109-122.

Grantham, R. (1980). Workings of the genetic code. *TIBS* **12**:327-331.

Green, M.R. (1986). Pre-mRNA splicing. *Annual Review of Genetics* **20:** 671-708.

Gunderson, J.H., Elwood, H., Ingold, A., Kindle, K., and Sogin, M.L. (1987). Phylogenetic relationships between chlorophytes, chrysophytes and oomycetes. *Proceedings of the National Academy of Sciences* USA **84:** 5823-5827.

Gupta, R.S., Aitken, K., Falah, M., and Singh, B. (1994). Cloning of *Giardia lamblia* heat shock protein HSP70 homologs: implications regarding origin of eukaryotic cells and of endoplasmic reticulum. *Proceedings of the National Academy of Sciences USA* **91:** 2895-2899.

Hanley, B.A., and Schuler, M.A. (1988). Plant intron sequences: evidence for distinct groups of introns. *Nucleic Acids Research* **16(14):** 7159-7177.

Harris, S., Rudnicki, K.S., and Haber, J.E. (1993). Gene conversions and crossing over during homologous and homeologous ectopic recombination *in Saccharomyces cerevisiae. Genetics* **135**: 5-16.

Hasabe, M., Ito, M., Kofuji, R., Ueda, K., and Iwatsuki, K. (1993). Phylogenetic relationships of ferns deduced from rbcL gene sequence. *Journal of Molecular Evolution* **37**:476-482.

Hasegawa, M., Hashimoto, T., Adachi, J., Iwabe, N., and Miyata, T. (1993). Early branchings in the evolution of eukaryotes: ancient divergence of entamoeba that lacks mitochondria revealed by protein sequence data. *Journal of Molecular Evolution* **36:** 380-388.

Hashimoto, T., Nakamura, Y., Nakamura, F., Shirakura, T., Adachi, J., Goto, N., Okamoto, K., and Hasegawa, M. (1994). Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia. Molecular Biology and Evolution* **11(1):** 65-71.

Helentjaris, T. (1993). Implications for conserved genomic structure among plant species. *Proceedings of the National Academy of Sciences, USA* **90:** 8308-8309.

Hendriks, L., Goris, A., Neefs, J-M., De Peer, Y.V., Hennebert, G., and De Wachter, R. (1989). The nucleotide sequence of the small ribosomal subunit RNA of the *yeast Candida albicans* and the evolutionary position of the fungi among the eukaryotes. *Systematics and Applied Microbiology* **12:** 223-229.

Hickey, D.A. (1992). Evolutionary dynamics of transposible elements in prokaryotes and eukaryotes. *Genetica*----------

Hightower, R.C., and Meagher, R.B. (1985). Divergence and differential expression of soybean actin genes. *EMBO J*, **4:**1-4.

Hightower, R.C., and Meagher, R.B. (1986). The molecular evolution of actin. *Genetics*, **114**:315-332.

Hillis, D.M., and Bull, J.J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biologist* **42(2):** 182-192.

Holland, S.K., and Blake, C.C.F., (1990). Proteins, exons, and molecular evolution. In: *Intervening sequences in evolution and development.* Eds, Stone, E.M., and R.J.Schwartz. Oxford University Press.pp 10-42.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6:** 65-70.

Holmes, K.C., Popp, D., Gerbhard, W., and Kabsch, W. (1990). Atomic model of the actin filament. *Nature* **347**: 44-49.

Hosaka, K. (1986). Who is the mother of the potato?- restriction endonuclease analysis of chloroplast DNA of cultivated potatoes. *Theoredical and Applied Genetics* **72:** 606-618.

Huynen, M.A., Konings, D.A.M., and Hogeweg, P. (1992). Equal G and C contents in histone genes indicate selection pressures on mRNA secondary structure. *Journal of Molecular Evolution* **34**: 280-291.

Kabsch, W., Mannherz, H.G., Suck, D., Pai, E.F., and Holmes, K.C. (1990). Atomic structure of the actin:DNase I complex. *Nature* **347:** 37-44.

Kawamoto, T., Makino, K., Niwa, H., Sugiyama, H., Kimura, S., Amemura, M., Nakata, A., and Kagunaga, T. (1988). Identification of the human β-actin enhancer and its binding factor. *Molecular and Cellular Biology* **8:** 267-272.

Kenny, J.R., Dancik, B.P., Florence, L.Z., and Nargang, F.E. (1988). Necleotide sequence of the carboxy-terminal of a lodgepole pine actin gene. *Canadian Journal of Forestry Res.* **18:** 1595-1602.

Kovacs, A.M., and Zimmer, W.E. (1993). Molecular cloning and characterization of the chicken smooth γ-actin mRNA. *Cell Motility and the Cytoskeleton* **24**: 67-81.

Kovilur, S., Jacobson, J.W., Beach, R.L., Jeffery, W.R., and Tomlinson, C.R. (1993). The evolution of the chordate muscle actin gene. *Journal of Molecular Evolution* **36:** 361-368.

Kumar, S., Tamura, K., and Nei, M., (1993). *MEGA*: Molecular evolutionary genetic analysis, version 1.0, The Pennsylvania State University, University Park, PA 16802.

Li, W-H., and Tanimura, M. (1987). The molecular clock runs more slowly in man than in apes and monkeys. *Nature* **326**: 93-96.

Li, W-H., (1983). Evolution of duplicate genes and pseudogenes. In: *Evolution of genes and proteins*, eds M. Nei and R.K. Koehn. Sinauer Associates, Inc.pp14-36.

Li, W-H., (1993a). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol,* **36:**96-99.

Li, W-H., (1993b). So, what about the molecular clock hypothesis? *Current Opinion in Genetics and Development* **3:**896-901.

Li, W.-H., Wu, C.-I., & Luo, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelyhood of nucleotide and codon changes. *Molecular Biology and Evolution*, **2**:150-174.

Li, W-H., and Graur, D. (1991). Fundamentals of molecular evolution. Sinauer Associates, Inc.

Lloyd, A.T., and Sharp, P.M. (1991). Codon usage in *Aspergillus nidulans. Molecular and General Genetics* **230:** 288-294.

Lloyd, A.T., and Sharp, P.M. (1992). CODONS: a microcomputer program for codon usage analysis. *J. Hered.*, **83**:239-240.

Lockhart, P.J., Howe, C.J., Bryant, D.A., Beanland, T.J., and Larkhum, A.W.D. (1992a). Substitution bias confounds inference of cyanelle origins from sequence data. *Journal of Molecular Evolution* **34:** 153-162.

Lockhart, P.J., Penny, D., Hendy, M.D., Howe, C.J., Beanland, T.J., and Larkhum, A.W.D. (1992b). Controversy on chloroplast origins. *FEBS* **301(2):** 127-131.

Loomis, W.F., and Smith, D.W. (1990). Molecular phylogegy of *Dyctyostelium discoideum* by protein sequence comparison. *Proceedings of the National Academy of Sciences*, USA **87:** 9093-9097.

Maniatis, T., Frisch, E.F., and Sambrock, J. (1982). Molecular Cloning. A Laboratory Manual, Cold Spring Harbor Laboratory, New York.

Margulis, L., and Schwartz, K.V. (1988). Five kingdoms: an illustrated guide to the phyla of life on earth. Second Edition, W.H. Freeman and Company, New York.

Martin, W., Gierl, A., and Saedler, H. (1989). Molecular evidence for pre-Cretacious angiosperm origins. *Nature* **339**: 46-48

Martin, P.G., Dowd, J.M., Morris, C., and Symon, D.E. (1986). The study of plant phylogeny using amino acid sequences of ribulose-1,5-biphosphate carboxylase. VI some *Solanum* and allied species from different continents. *Australian Journal of Botany* **34:** 187-195.

McElroy, D., Rothenberg, M., and Wu, R. (1990). Structural characterization of a rice actin gene. *Plant Molecular Biology* **14:** 163-171.

McHugh, K.M., Crawford, K., and Lessard, J.L. (1991). A comprehensive analysis of the developmental and tissue-specific expression of the isoactin multigene family in the rat. *Developmental Biology* **148**: 442-458.

McLean, M., Baird, W.V., Gerats, A.G.M., and Meagher, R.B. (1988). Determination of copy number and linkage relationships among five actin gene subfamilies in *Petunia hybrida. Plant Molecular Biology* **11**: 663-672.

McLean, M., Gerats, A.G.M., Baird, W.V., and Meagher, R.B. (1990). Six actin gene subfamilies map to five chromosomes of *Petunia hybrida. Journal of Heredity* **81**:341-346.

Meagher, R.B. (1991). Divergence and differential expression of actin gene families in higher plants. *International Review of Cytology* **125**:139-163.

Meagher, R.B., Berry-Lowe, S., & Rice, K. (1989). Molecular evolution of the small subunit of ribulose bisphosphate carboxylase: nucleotide substitution and gene conversion. *Genetics* **123**:845-863.

Melamed, C., Nevo, Y., and Kupiec, M. (1992). Involvement of cDNA in homologous recombination between Ty elements *in Saccharomyces cereviciae. Molecular and Cellular Biology* **12(4):** 1613-1620.

Mouchiroud, D., and Bernardi, G. (1993). Compositional properties of coding sequences and mammalian phylogeny. *Journal of Molecular Evolution* **37:** 109-116.

Murray, E.E., Lotzer, J., and Eberle, M. (1989). Codon usage in plant genes. *Nucleic Acids Research* **17(2):** 477-498.

Nairn, J.C., Winesett, L., and Ferl, R.J. (1988). Nucleotide sequence of an actin gene from *Arabidopsis thaliana. Gene* **65**:247-257.

Ohta, T. (1991). Multigene families and the evolution of complexity. *Journal of Molecular Evolution* **33**:34-41

Palmer, J.D., and Logsdon, J.M. (1991). The recent origins of introns. *Current Opinion in Genetics and Development* **1**:470-477.

Parsons, T.J., Bradshaw, H.D., and Gordon, M.P. (1989). Systemic accumulation of specific mRNAs in response to wounding of poplar trees. *Proceedings of the National Academy of Sciences* USA **86:** 7895-7899.

Patterson, C. (1989). Phylogenetic relations of major groups: conclusions and prospects. In: B. Fernholm, K. Bremer, & H. Jörnvall (eds), *The Hierarchy of Life*. Elsevier: Amsterdam, 471-489.

Pearson, W.R. (1994). The FASTA program package. University of Virginia.

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

Pearson, L., and Meagher, R.B. (1990). Diverse soybean actin transcripts contain a large intron in the 5' untranslated leader: structural similarity to vertebrate muscle actin genes. *Plant Molecular Biology* **14**: 513-526.

Perlman, P.S., Peebles, C.L., and Daniels, C. (1990). Different types of introns and splicing mechanisms. In: *Intervening sequences in evolution and development.* Eds, Stone, E.M. and R.J. Schwartz. Oxford University Press. pp 112-161.

Reece, K.S., McElroy, D., and Wu, R. (1992). Function and evolution of actins. *Evolutionary Biology* **26**:1-34.

Rice, W.R. (1989). Analyzing tables of statistical tests. *Evolution* **43(1):** 223-225.

Ricklefs, E. (1990). Ecology. Third Edition. W.H. Freeman and Company. p287.

Saitou, N., and Imanishi, T. (1989). Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelyhood, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. *Molecular Biology and Evolution* **6(5):** 514-525.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4:**406-425.

Sanger, F., Nicklen, S., & Coulson, A.R. (1977). DNA sequencing with chain-termination inhibitors. *Proceedings of the National Academy of Sciences, USA,* **74**:5463-5467.

Sarich, V.M., and Wilson, A.C. (1973). Generation time and genomic evolution in primates. *Science* **179**: 1144-1147.

Savard, L., Li, P., Strauss, S.H., Chase, M.W., Michaud, M., and Bousquet, J. (1994). Chloroplast and nuclear gene sequences indicate late Pennsylvanian time for the last common ancestor of extant seed plants. *Proceedings of the National Academy of Sciences, USA* **91**: 5163-5167.

Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Molecular Biology and Evolution* **6(5):** 526-538.

Schlegel, M. (1994). Molecular phylogeny of eukaryotes. *Trends in Ecology and Evolution* **9(9):** 330-335.

Schwob, E., and Martin, R.P. (1992). New yeast actin-like gene required late in the cell cycle. *Nature* **355**: 179-182.

Seagull, R.W. (1989). The plant cytoskeleton. *Critical Reviews In Plant Sciences* **8(2)**:131-167.

Shah, D.M., Hightower, R.C., and Meagher, R.B. (1982). Complete nucleotide sequence of a soybean actin gene. *Proc. Natl. Acad. Sci. USA*, **79**:1022-1026.

Sharp, P.M., and Li, W.-H. (1986). An evolutionary perspective on synonomous codon usage in unicellular organisms. *Journal of Molecular Evolution* **24**: 28-38.

Sheterline, P., and Sparrow, J.C. (1994). Actin. *Protein Profile* **1(1)**: 1-121.

Sidow, A., and Thomas, W.K. (1994). A molecular evolutionary framework for eukaryotic model organisms. *Current Biology* **4(7):** 595-603.

Sogin, M.L. (1989). Evolution of eukaryotic microorganisms and their small subunit ribosomal RNAs. *American Zoologist* **29:** 487-499.

Sogin, M.L. (1991). Early evolution and the origin of eukaryotes. *Current opinion in Genetics and Development* **1:** 457-463.

Soltis, P.S., Doyle, J.J., and Soltis, D.E. (1992). Molecular data and polyploid evolution in plants. In: P.S. Soltis, D.E. Soltis, and J.J. Doyle (Eds) *Molecular systematics of plants*. Chapman and Hall.

Stewart. W.N. (1983). Paleobotany and the evolution of plants. Cambridge University Press, Cambridge. pp. 397.

Stoltzfus, A., Spencer, D.F., Zuker, M., Logsdon, J.M., and Doolittle, W.F. (1994). Testing the exon theory of genes: the evidence from protein structure. *Science* **265**: 202-207.

Stranathan, M., Hastings, C., Trinh, H., and Zimmerman, J.L. (1989). Molecular evolution of two actin genes from carrot. *Plant Molecular Biology* **13**:375-383.

Tanksley, S.D. (1987). Organization of the nuclear genome in tomato and related diploid species. *The American Naturalist* **130:** S46-S56.

Thangavelu, M., Belostotsky, D., Bevan, M.W., Flavell, R.B., Rogers, H.J., and Lonsdale, D.M. (1993). Partial characterization of the *Nicotiana tabacum* actin gene family: evidence from pollen specific expression of one of the gene family members. *Molecular and General Genetics* **240:** 290-295.

Vandekerckhove, J., and Weber, K. (1984). Chordate muscle actins differ distinctly from invertebrate muscle actins. *Journal of Molecular Biology* **179:** 391-413.

Vanin, E.F. (1985). Processed pseudogenes: characteristics and evolution. *Annual Review of Genetics* **19**: 253-272.

Wainright, P.O., Hinkle, G., Sogin, M.L., and Stickel, S.K. (1993). Monophyletic origins of the metazoa: an evolutionary link with fungi. *Science* **260:** 340-342.

Weber, K., and Kabsch, W. (1994). Intron positions in actin genes seem unrelated to the secondary structure of the protein.*The EMBO Journal* **13(6)** : 1280-1286.

Wells, D., Bains, W., and Kedes, L. (1986). Codon usage in histone gene families of higher eukaryotes reflects functional rather than phylogenetic relationships. *Journal of Molecular Evolution* **23:** 224-241.

White, O., Soderlund, C., Shanmugan, P., and Fields, C. (1992). Information contents and dinucleotide compositions of plant intron sequences vary with evolutionary origin. *Plant Molecular Biology* **19**: 1057-1064.

Wilbur, J.W. (1985). On the PAM matrix model of protein evolution. *Molecular Biology and Evolution* **2(5):** 434-447.

Wildeman, A.G. (1988). A putative ancestral actin gene present in a thermophilic eukaryote: novel combination of intron positions. *Nucleic Acids Research* **16(6):** 2553:2564.

Wolf, S.J., and Timko, M.P. (1994). Characterization of actin-gene family members and their expression during developement in witchweed (*Striga asiatica* L.). *Planta* **192**: 61-68.

Wolfe, K.H., Li, W-H., and Sharp, P.M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Science, USA* **84**:9054-9058.

Wolfe, K.H., Gouy, M., Yang, Y-W., Sharp, P.M., and Li, W-H. (1989). Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proceedings of the National Academy of Science, USA* **86:**6201-6205.

Yoshida, E.N., and Hickey, D.A. (1995). Nucleotide bias can affect protein evolution. *Unpublished.-*

Zar, J.H. (1984). Biostatistical analysis. Second edition. Prentice-Hall, Inc.

```
                                                                                                                    70
vlvx       GCTGGCTTTG CTGGTGATGA TGCCCCACGA GCTGTGTTTC CGAGCATTGT TGGTCGGCCC CGCCATACGG
PsAc2      GCTGGATTTG CTGGTGATGA TGCT---AGG GCTGTGTTTC CAAGTATTGT TGGCCGACCT CGTCATACTG
PsAc1      GCTGGATTTG CTGGTGATGA TGCT---AGG GCTGTGTTTC CAAGTATTGT TGGCCGACCT CGTCATACTG
SoAc1      GCTGGGTTCG CTGGAGATGA CGCCCCCAGG GCCGTCTTCC CCAGCATTGT CGGCCGGCCG CGCCACACCG
AAc1       GCTGGTTTTG CTGGGGATGA TGCACCTAGA GCTGTGTTTC CTAGTATTGT GGGTCGTCCT CGTCACACCG
RAc1       GCTGGGTTCG CCGGAGATGA TGCGCCCAGG GCTGTCTTCC CCAGCATTGT CGGCCGCCCT CGCCACACCG
RAc3       GCTGGTTTTG CTGGTGATGA TGCCCCCAGA GCAGTGTTCC CTAGTATAGT TGGTCAGCCT CGGCACACTG
RAc7       GCCGGTTTTG CTGGTGATGA TGATGCACCC AGGGCTGTCT TTCCTAGCAT TGTAGGCAGG CCACGCCACA
RAc2       GCTGGTTTGG CTGGAGATGA TGCTCCCCGT GCTGTTTTCC CGAGTATTGT TGGCCGTCCG CGACATACAG
DcAc1      GCTGGATTTG CTGGTGATGA TGCTCCAAGC GCTGTGTTTC CCAGTATTGT T---GGCAGA CCTCGACACA
pt46       GCAGGATTTG CAGGAGATGA TGCACCTAGG GCTGTTTTTC CTAGCATTGT TGGGCGTCCG CGTCATGCTG
pt66       GCTGGATTTG CTGGAGATGA TGCTCCAAGG GCTGTCTTCC CTAGTATTGT GGGACGTCCT CGTCACACTG
pt79       GCTGGATTTG CTGGAGATGA TGCTCCAAGG GCTGTCTTCC CTAGCATTGT GGGACGTCCT CGTCACACTG
pt82       GCGGGATTTG CTGGAGATGA TGCTCCAAGA GCAGTATTTC CTAGCATTGT TGGTCGCCCT CGTCATACAG
PoAc101 GCTGGATTTG CTGGTGATGA TGCACCAAGG GCTGTATTTC CTAGTATTGT CGGCCGACCT AGACACACCG
PoAc58     GCTGGGTTTG CGGGAGATGA TGCTCCACGA GCTGTATTCC CACGTATTGT TGGTCGTCCG CGTCATACTG
PoAc71     GCAGGATTTG CTGGAGATGA TGCACCTAGG GCTGTTTTTC CTAGCATTGT TGGGCGTCCG CGTCATGCTG
PoAc75     GCTGGTTTTG CTGGAGATGA CTCCCCAAGG GCAGTTTTTC CTAGCATTGT TGGACGTCCT CGTCACACTG
PoAc97     GCTGGGTTTG CTGGAGATGA TGCTCCACGA GCTGTATTTC CTAGTATTGT TGGCCGTCCC CGCCATAGTG
PoAc99     GCTGGATTTG CTGGTGATGA TGCTCCTAGA GTTGTGTTCA CTAGCATTGT TGGTCGTCCT CGCCACACTG
pt42       GCTGGATTTG CTGGAGACGA TGCTCCACGA GCAGTGTTTC CGAGCATTGT TGGTCGGCCA CGACACACAG
pt65       GCTGGATTTG CTGGAGACGA TGCTCCACGA GCAGTGTTTC CGAGCATTGT TGGTCGGCCA CGACACACAG
tm32       GCTGGATTTG CTGGAGACGA TGCGCCACGA GCAGTGTTTC CTAGCATTGT TGGTAGACCA CGTCACACGG
tm41       GCTGGGTTCG CAGGAGATGA TGCTCCACGA GCTGTATTTC CTAGTATTGT TGGCCGCCCC CGCCATACTG
tm52       GCTGGATTTG CTGGAGATGA TGCTCCAAGG GCTGTCTTCC CTAGTATTGT GGGACGTCCT CGTCACACTG
tm51       GCTGGATTTG CTGGTGATGA TGCCCCTAGA GCAGTGTTTC CTAGCATTGT TGGTCGTCCT CGCCACACTG
tm105      GCCGGTTTTG CTGGGGATGA TGCTCCAAGA GCTGTTTTTC CAAGTATAGT TGGTCGTCCT CGTTACACAG
tb103      GCTGGATTTG CTGGTGATGA TGCTCCTAGA GCTGTGTTTC CTAGCATTGT GGGTCATCCT CGACACACTG
tb104      GCTGGCTTTG CTGGAGATGA TGCTCCAAGG GCTGTATTTC CTAGTATAGT TGGTCGTCCT AGACACCAAG
tb66       GCTGGATTTG CTGGTGATGA TGCTCCGAGG GCTGTTTTTC CTAGTATTGT TGGTCGACCT CGACACACTG
tb71       GCTGGGTTTG CTGGAGATGA TGCTCCACGA GCTGTATTCC CAAGTATTGT TGGGCAGACC CGCCATACTG
tb93       GCTGGATTCG CCGGAGATGA TGCTCCAAGA GCTGTGTTCC CTAGCATTGT TGGACGCCCT CGTCGCACTG
tob1       GCAGGGTTTG CTGGAGATGA TGCTCCAAGA GCAGTATTTC CGAGTATTGT TGGTCGCCCA CGCCACACAG
tb53       GCTGGTTTTG CTGGGGATGA TGCTCCAAGG GCTGTTTTTC CAAGTATAGT TGGTCGTCCT CGTTACACTG
tb54       GCTGGATTTG CTGGTGATGA TGCTCCTCGA GCAGTATTTC CCAGCATTGT GGGTCGCCCT CGGCACACTG
soy109     GCTGGATTTG CTGGTGATGA TGCTCCCAGG GCTGTCTTTC CCAGTATTGT TGGCCGACCT CGACACACTG
soy115     GCTGGATTTG CTGGGGATGA TGCTCCAAGG GCTGTGTTTC CCAGCATTGT TGGTCGTCCC CGTCACACGG
soy118     GCTGGTTTTG CTGGAGATGA TGCCCCACGT GCTGTGTTTC CCAGCATTGT AGGTCGTCCT CGTCACACTG
soy119     GCTGGTTTTG CTGGAGATGA TGCTCCGAGG GCAGTGTTTC CAAGCATTGT TGGTCGGCCA CGTCACACCG
soy57      GCTGGTTTTG CAGGTGATGA TGCTCCTAGG GCCGTGTTCC CCAGCATTAT TGGGAGACCA CGCCATACTG
soy58      GCAGGATTTG CTGGTGACGA TGCTCCTAGG GCTGTGTTTC CCAGTATTGT TGGTCGACCT CGTCATACTG
soy69      GCAGGATTTG CAGGAGACGA TGCCCCACGT GCAGTGTTTC CTAGCATTGT TGGTCGTCCT CGTCTCACTG
soy70      GCTGGATTTG CTGGGGATGA TGCTCCGAGG GCTGTGTTTC CCAGCATTGT TGGTCGTCCC CGTCACACGG
soy86      GCAGGATTTG CTGGTGACGA TGCTCCTAGG GCTGTCTTTC CCAGTATTGT TGGCCGACCT CGACATACTG
SAc1       GCAGGACTTG CAGGAGATGA TGCCCCACGT GCTGTGTTTC CTAGCATTGT TGGTCGTCCT CGTCACACT?
SAc3       GCAGGATTTG CTGGTGACGA TGCTCCTAGG GCTGTCTTTC CCAGTATTGT TGGCCGACCT CGACATACTG
mz56       GCAGGTTTCG CTGGTGACGA CGCGCCGAGG GCAGTCTTTC CCAGCATCGT AGGCAGGCCA CGGCACACTG
mz63       GCAGGTTTCG CTGGTGACGA CGCGCCGAGG GCAGTCTTTC CCAGCATCGT AGGCAGGCCA CGGCACACTG
mz81       GCTGGTTTTG CTGGTGACGA TGCGCCAAGG GCTGTTTTCC CAAGCATTGT TGGACGCCCA CGTCACACTG
mz83       GCTGGTTTTG CTGGTGACGA TGCGCCAAGG GCTGTTTTCC CAAGCATTGT CGGACGCCCC CGTCACACTG
mz87       GCTGGTTTTG CCGGCGATGA TGCACCAAGG GCTGTATTCC CTAGCATCGT TGGGCGTCCT CGTCACACCG
mz89       GCGGGTTTCG CTGGTGACGA CGCACCGAGA GCTGTCTTTC CTAGCATTGT AGGCAGGCCA CGCCACACTG
mz95       GCTGGCTTTG CTGGTGATGA TGCGCCCAGA GCAGTGTTCC CTAGTATAGT CGGCCGCCCT CCGGACACGG
MAc1       GCCGGTTTCG CTGGTGATGA TGCGCCAAGA GCTGTCTTCC CCAGCATTGT GGGAAGACCA CGCCACACCG
```

Appendix: Alignment of 54 plant actin nucleotide sequences used to generate phylogenetic trees.

```
                                                                                    140
vlvx     GT---GTGAT GGTCGGCATG GGGCAGAAGG ATTCCTACGT GGGCGACGAG GCACAGTCTA AGCGTGGT--
PoAc2    GT---GTCAT GGTTGGGATG GGTCAAAAGG ATGCCTATGT AGGTGATGAA GCCCAATCTA AAAGAGGT--
PoAc1    GT---GTCAT GGTTGGGATG GGTCAAAAGG ATGCCTATGT AGGTGATGAA GCCCAATCTA AAAGAGGT--
SoAc1    GT---GTCAT GGTCGGGATG GGGCAGAAGG ACGCCTACGT TGGTGACGAG GCGCAGTCCA AGAGGGGT--
AAc1     GT---GTGAT GGTTGGGATG GGGCAAAAGG ATGCTTATGT TGGCGATGAA GCTCAATCCA AACGAGGT--
RAc1     GT---GTCAT GGTCGGAATG GGCCAGAAGG ACGCCTACGT CGGCGACGAG GCGCAGTCCA AGAGGGGT--
RAc3     GT---GTCAT GGTTGGAATG GGCCAAAAAG ATGCCTATGT GGGTGATGAG GCACAATCAA AAAGAGGT--
RAc7     GCGGTGTCAT GGTTGGTATG GGCCAGAAGG ATGCCTATGT GGGTGATGAA GCTGCAGTCG AAAAGAGGGA
RAc2     GC---GTTAT GGTTGGGATG GGACAAAAAG ATGCTTATGT CGGTGATGAG GCCCAATCCA AGAGGGGT--
DcAc1    CTGGTGTGAT GGTTGGAATG GGCCAAAAGG ATGCCTATGT TGGTGATGAA GCACAATCCA AAAGGGGT--
pt46     GA---GTGAT GGTAGGAATG GGACAGAAGG ATGCATATGT TGGCGATGAG GCACAGTCGA AACGTGGT--
pt66     GT---GTTAT GGTTGGCATG GGCCAGAAAG ATGCCTATGT TGGTGATGAG GCTCAATCCA AAAGGGGT--
pt79     GT---GTGAT GGTTGGCATG GGCCAGAAAG ATGCCTATGT TGGTGATGAG GCTCAATCCA AAAGGGGT--
pt82     GA---GTAAT GGTAGGAATG GGGCAAAAAG ATGCATATGT TGGAGATGAA GCTCAATCCA AACGTGGT--
PoAc101  GT---GTTAT GGTAGGGATG GGTCAAAAGG ATGCATATGT TGGTGATGAA GCTCAGTCAA AAAGGGGT--
PoAc58   GT---GTGAT GGTTGGTATG GGTCAAAAAG ATGCTTATGT AGGAGATGAA GCTCAATCAA AGCGAGGT--
PoAc71   GA---GTGAT GGTAGGAATG GGACAGAAGG ATGCATATGT TGTTGATGAG GCTCAGTCGA AACGTGGT--
PoAc75   GA---GTGAT GGTTGGAATG GGCCAGAAGG ATGCCTATGT TGGAGATGAG GCTCAGTCGA AACGTGGT--
PoAc97   GT---GTGAT GGTGGGTATG GGTCAAAAAG ACGCCTATGT GGGAGATGAA GCTCAATCGA AGAGAGGT--
PoAc99   GT---GTTAT GGTTGGAATG ACACAGAAGG ATGCATATGT ?GGTGATGAA GCTCAATCCA AGAGGGGT--
pt42     GT---GTGAT GGTTGGTATG GGGCAAAAAG ATGCATATGT TGGTGATGAA GCTCAATCGA AACGTGGT--
pt65     GT---GTGAT GGTTGGTATG GGGCAAAAAG ATGCATATGT TGGTGATGAA GCTCAATCGA AACGTGGT--
tm32     GT---GTTAT GGTTGGTATG GGTCAAAAAG ATGCGTATGT TGGTGATGAA GCTCAATCG? AACGTGGT--
tm41     GT---GTGAT GGTGGGTATG GGTCAAAAAG ACGCCTATGT GGGAGATGAA GCTCAATCGA AGAGAGGT--
tm52     GT---GTGAT GGTTGGCATG GGCCAGAAAG ATGCCTATGT TGGTGACGAG GCTCAATCCA AAAGGGGT--
tm51     GT---GTTAT GGTTGGAATG GGTCAGAAAG ATGCATATGA TGGTGATGAA GCTCAATCCA AGAGGGGT--
tm105    GT---GTCAT GGTTGGAATG GGACAAAAAG ATGCCTATGT TGGTGATGAA GCTCAGTCCA AAAGAGGT--
tb103    GC---GTTAT GGTTGGAATG GGACAGAAAG ATGCCTATGT GGGTGACGAA GCTCAATCCA AGAGGGGT--
tb104    GA---GTAAT GGTTGGGATG GGGCAAAAAG ATGCCTATGT TGGTGATGAA GCTCAGTCTA AAAGAGGT--
tb66     GT---GTTAT GGTTGGTATG GGTCAAAAGG ATGCATATAT TGGTGACGAA GCTCAGTCCA AAAGGGGT--
tb71     GT---GTGAT GGTGGGTATG GGT???AAAG ATGCCTACGT GGGAGATGAA GCTCAATCAA AAAGAGGT--
tb93     GT---GTGAT GGTTGGCATG GGCCAGAAGG ATGCTTATGT TGGTGATGAG GCTCAATCCA AAAGGGGT--
tob1     GT---GTGAT GGTTGGCATG GGGCAAAAAG ATGCATATGT TGGTGATGAA GCTCAATCCA AACGTGGT--
tb53     GC---GTCAT GGTTGGAATG GGACAAAAAG ATGCCTATGT TGGTGATGAA GCTCAGTCTA AAAGAGGT--
tb54     GT---GTTAT GGTCGGAATG GGACAGAAGG ATGCTTATGT GGGTGATGAA GCTCAATCCA AGAGAGGT--
soy109   GT---GTTAT GGTTGGGATG GGTCAAAAGG ATGCCTATGT TGGTGATGAA GCCCAATCAA AAAGAGGT--
soy115   GA---GTGAT GGTTGGCATG GGCCTAAAAG ATGCATATGT CGGGGATGAA GCTCAGTCCA AGAGAGGT--
soy118   GT---GTGAT GGTTGGCATG GGCCAAAAAG ATGCATATGT TGGTGATGAG GCACAGTCTA AGAGGGGT--
soy119   GT---GTGAT GGTTGGGATG GGGCAAAAGG ATGCATATGT TGGGGATGAG GCTCAGTCCA AGCGTGGT--
soy57    GT---GTAAT GGTTGGGATG GGTCAGAAGG ATGCTTACGT GGGTGATGAA GCCCAATCAA AAAGGGGT--
soy58    GT---GTTAT GGTTGGGATG GGTCAAAAGG ATGCCTATGT AGGTGATGAA GCCCAGTCAA AAAGAGGT--
soy69    GT---GTGAT GGTTGGTATG GGCCAGAAAG ATGCATATGT TGGTGATGAG GCACAATCTA AACGTGGT--
soy70    GA---GTGAT GGTTGGCATG GGCCAAAAAG ATGCATATGT TGGGGATGAA GCTCAATCCA AGAGAGGT--
soy86    GT---GTTAT GGTTGGGATG GGTCAAAAGG ATGCCTATGT TGGTGATGAA GCCCAATCCA AAAGAGGT--
SAc1     ?A---GTTGT AGTTGGTATG GGCCAGAAAG ATGCATATGT TGGTGATGAA GCACAAGTCC AACGTGGT--
SAc3     GT---GTTAT GGTTGGGATG GGTCAAAAGG ATG?CTATGT TGGTGATGAA GCCCAATCCA AAAGAGGT--
mz56     GC---GTGAT GGTGGGCATG GGCCAAAAGG ATGCATACGT GGGCGACGAA GCCCAGTCCA AGAGAGGC--
mz63     GT---C??AT GGTGGGCATG GGCCAAAAGG ATGCATACGT GGGCGACGGA GCCCAGTCCA AGAGAGGC--
mz81     GC---GTGAT GGTTGGCATG GGGCAGAAGG ATGCGTATGT TGGCGATGAG GCTCAGTCCA AGAGGGGT--
mz83     GT---GTGAT GGTTGGCATG GGGCAGAAGG ATGCGTATGT TGGTGACGAG GCCCAGTCCA AGAGGGGT--
mz87     GT---GTCAT GGTAGGGATG GGGCAGAAGG ATGCATATGT TGGTGATGAG GCACAGTCCA AGAGAGGT--
mz89     GT---GTGAT GGTGGGCATG GGCCAAAAGG ATGCATATGT GGGCGATGAA GCTCAGTCCA AAAGAGGT--
mz95     GT---GTGAT GGTTGGAATG GGCCAGAAGG ACGCCTGTGT AGGTGACGAG GCACAGTCAA AAAGAGGT--
MAc1     GT---GTCAT GGTCGGCATG GGCCAAAAGG ATGCCTACGT AGGTGATGAG GCTCAGGCCA AGAGAGGC--
```

Appendix: con't

```
                                                                                                                    210
v1vx       -ATCTTGACG TTGCGC---T ACCCTATTGA GCACGGTATC GTTACCAACT GGGATGACAT GGAGAAGATC
PsAc2      -ATTCTTACT CTTAAG---T ACCCCATTGA GCACGGTATA GTCAGTAACT GGGATGACAT GGAGAAGATC
PsAc1      -ATTCTTACT CTTAAG---T ACCCCATTGA GCACGGTATA GTCAGTAACT GGGATGACAT GGAGAAGATC
SoAc1      -ATCCTGACC CTCAAG---T ACCCCATCGA GCACGGAATC GTCAGCAACT GGGACGATAT GGAGAAGATC
AAc1       -ATTTTAACT CTCAAG---T ACCCTATTGA GCATGGAATT GTCAACAATT GGGATGACAT GGAGAAGATT
RAc1       -ATCTTGACC CTCAAG---T ACCCCATCGA GCATGGTATC GTCAGCAACT GGGATGATAT GGAGAAGATC
RAc3       -ATCTTGACC TTGAAA---T ACCCTATTGA GCATGGTATG TGGAACAACT GGGACGACAT GGAGAAAATC
RAc7       TATCTTGACA TTGAAA---T ACCCAATCGA GCATGGTATT GTCAACAACT GGGATGACAT GGAGAAGATA
RAc2       -ATCCTAACC TTGAAA---T ACCCCATTGA GCATGGAATT GTAAGCAACT GGGATGACAT GGAGAAAATT
DcAc1      -ATCATCACA TTGAAA---T ACCCAATTGA ACACGGCATT GTTAGCAATT GGGATGACAT GAGA---ATA
pt46       -ATTTTGACA CTGAAG---T ATCCAATTGA ACATGGCATA GTGAATAATT GGGATGACAT GGAGAAGATT
pt66       -ATTTTGACT TTGAAG---T ATCCGATTGA ACACGGTATT GTTAGCAATT GGGATGACAT GGAGAAGATT
pt79       -ATTTTGACT TTGAAG---T ATCCGATTGA GCACGGTATT GTTAGCAATT GGGATGACAT GGAGAAGATT
pt82       -ATTCTCACA TTGAAA---T ACCCAATTGA GCACGGTATT GTTAGCAATT GGGATGATAT GGAGAAAATT
PoAc101    -ATTCTCACC TTAAAG---T ATCCCATTGA GCATGGAATA GTAAGCAACT GGGATGATAT GGAGAAGATT
PoAc58     -ATCTTAACG TTGAAG---T ACCCGATTGA GCATGGTATT GTCAGCAATT GGGATGATAT GGAGAAGATA
PoAc71     -ATTCTTACA CTAAAG---T ATCCAA?TGA ACATGGCATA GTGAATAATT GGGATGACAT GGAGAAGATT
PoAc75     -ATTCTGACA TGGAAG---T ATCCTATTGA ACATGGTATT GTTAATAACT GGGATGACAT GGAGAAGATT
PoAc97     -ATTTTAACT CTTAAA---T ACCCAATTGA ACACGGAATT GTCAGCAATT GGGATGATAT GGAGAAGATA
PoAc99     -ATCTTGACC TTGAAAA??T AGCCAAGTGA GCATGGTATA GTCAGCAACT GGGATGATAT GGAGAAGATA
pt42       -ATTCTATCA TTGAAA---T ATCCAATTGA GCATGGTATT GTGAGCAATT GGGATGACAT GGAGAAAATC
pt65       -ATTCTATCA TTGAAA---T ATCCAATTGA GCATGGTATT GTGAGCAATT GGGATGACAT GGAGAAAATC
tm32       -ATTCTATCT TTGAAA---T ATCCAATTGA GCATGGGATT GTAAGTAATT GGGACGATAT GGAGAAAATA
tm41       -ATTTTAACT CTTAAA---T ACCCAATTGA GCACGGAATT GTCAGCAATT GGGATGATAT GGAGAAGATA
tm52       -ATTTTGACT TTGAAG---T ATCCGATTGA ACATGGTATT GTTAGCAATT GGGATGACAT GGAGAAGATT
tm51       -ATCTTGACC TTGAAA---T ACCCAATTGA ACACGGTATT GTCAGCAACT GGGATGATAT GGAGAAGATC
tm105      -ATTTTGACT CTGAAG---T ATCCAATTCG ACATGGTATA GTAGACAACT GGGATGATAT GGAGAAAATC
tb103      -ATCTTGACC TTGAAG---T ACCCAATTGA GCACGGTATA GTCAGCAACT GGGATGACAT GGAGAAGATA
tb104      -ATTCTGACT TTGAAA---T ATCCGATTGA ACATGGTATC GTCAGCAACT GGGATGACAT GGAGAAAATT
tb66       -ATTCTCACC TTGAAG---T ATCCCATTGA GCATGGAATA GTAAGCAACT GGGACGATAT GGAGAAGATC
tb71       -ATTTTAACT CTTAAA---T ACCCAATTGA GCATGGAATT GTCAGCAACT GGGATGATAT GGAGAAGATA
tb93       -ATTTTGACT TTGAAG---T ATCCAATTGA ACACGGAATC GTTAGCAATT GGGATGATAT GAAGAAAATT
tob1       -ATTCTAACT TTGAAA---T ATCCAATTGA ACATGGTATT GTTAGCAACT GGGATGATAT GGAGAAAATC
tb53       -ATTTTGACT CTAAAA---T ATCCGATTGA ACACGGCATA GTAAGCAACT GGGATGATAT GGAGAAAATA
tb54       -ATCTTGACC TTGAAA---T ATCCTATTGA GCATGGTATA GCCAGCAACT GGGATGATAT GGAGAAAATA
soy109     -ATTCTTACT CTCAAG---T ATCCCATTGA GCATGGTATA GTCAGCAACT GGGATGACAT GGAGAAGATC
soy115     -ATATTAACT CTCAAA---T ATCCCATTGA GCATGGTATT GTGAGTAACT GGGATGACAT GGAGAAGATC
soy118     -ATTTTGACT CTCAAG---T ACCCCATTGA ACACGGAATT GTGAGCAATT GGGATGATAT GGAAAAGATT
soy119     -ATACTAACT CTGAAA---T ATCCCATTGA GCATGGTATT GTCAGTAACT GGGATGACAT GGAGAAAATC
soy57      -ATACTCACC TTGAAG---T ATCCTATAGA GCATGGCATA GTCAGCAACT GGGACGATAT GGAGAAGATT
soy58      -ATTCTTACT TTGAAG---T ACCCAATCGA GCATGGTATA GTCAGTAACT GGGATGACAT GGAAAAGATC
soy69      -ATTCTGACT CTCAAG---T ATCCCATTGA GCATGGTATT GTTAGCAACT GGGATGATAT GGAAAAGATT
soy70      -ATATTAACT CTCAAA---T ATCCCATTGA GCATGGTATT GTGAGTAACT GGGATGACAT GGAGAAGATC
soy86      -ATTCTTACT CTCAAG---T ATCCCATTGA GCATGGTATA GTCAGCAACT GGGATGACAT GGAA?AGATC
SAc1       -ATTCTGACT CTCAAG---T ATCCGATTGA GCATGGTATT GTTAGCAACT GGGATGATAT GGAAAAGATT
SAc3       -ATTCTTACT CTCAAG---T ATCCCATTGA GCATGGTATA GTCAGCAACT GGGATGACAT GGAAAAGATC
mz56       -ATTCTGACG CTCAAG---T ACCCAATCGA GCACGGCATT GTGGGCAACT GGGACGACAT GGAGAAGATC
mz63       -ATTCTGACG CTCAAG---T ACCCAATCGA GCACGGCATT GTGGGCAACT GGGACGACAT GGAGAAGATC
mz81       -ATCCTGACT TTGAAG---T ACCCGATCGA ACATGGCATT GTCAACAACT GGGATGATAT GGAGAAGATC
mz83       -ATCCTGACT TTGAAG---T ACCCCATCGA ACATGGTATT GTTAGCAATT GGGATGACAT GGAGAAGATC
mz87       -ATTCTCACG CTCAAG---T ACCCGATTGA GCACGGTATT GTAAGCAACT GGGATGACAT GGAGAAAATC
mz89       -ATTCTGACA CTGAAG---T ACCCAATTGA GCACGGCATT GTCGGCAACT GGGATGACAT GGAGAAGATC
mz95       -ATCTTGACC TTGAAG---T ACCCCATTGA GCACGGTATC GTGAACAACT GGGATGATAT GGAGAAAATC
MAc1       -ATCCTGACA CTGAAG---T ACCCGATTGA GCATGGCATT GTCAACAACT GGGATGACAT GGAGAAC---
```

Appendix: con't

							280
vlvx	TGGCACCACA	CTTTCTTTAA	TGAGCTTCGT	GTGGCACCGG	AGGAGCACCC	AGTCCTTCTC	ACGGAGGCGC
PsAc2	TGGCACCACA	CATTTTACAA	TGAATTGCGT	GTTGCTCCTG	AGGAGCACCC	AGTGCTTCTA	ACTGAGGCTC
PsAc1	TGGCACCACA	CATTTTACAA	TGAATTGCGT	GTTGCTCCTG	AGGAGCACCC	AGTGCTTCTA	ACTGAGGCTC
SoAc1	TGGCATCACA	CCTTCTACAA	CGAGCTCCGT	GTGGCTCCCG	AGGAGCACCC	CGTCCTCCTC	ACTGAGGCGC
AAc1	TGGCATCACA	CTTTCTACAA	TGAGCTCCGT	GTTGCTCCTG	AGGAACATCC	TATTCTACTT	ACCGAGGCAC
RAc1	TGGCATCACA	CCTTCTACAA	CGAGCTCCGT	GTGGCCCCGG	AGGAGCACCC	CGTCCTCCTC	ACCGAGGCTC
RAc3	TGGCATCACA	CCTTCTACAA	TGAGCTGCGT	GTGCTCCCTG	AGGAACATCC	AATTTTGCTG	ACAGAAGCTC
RAc7	TGGCACCACA	CCTTCTACAA	TGAGCTCCGT	GTGGCTCCTG	AAGAGCATCC	TGTATTGCTG	ACCGAGGCTC
RAc2	TGGCACCACA	CATTCTACAA	TGAGCTTCGT	GTAGCACCAG	AAGAGCATCC	AATTCTTCTT	ACGGAGGCTC
DcAc1	TGGCATCACA	CTTTCTACAA	TGAGCTGCGT	GCTAGCCCTG	AAGAGCACCC	GGTACTACTT	ACAGAAGCAC
pt46	TGGCATCATA	CGTTCTATAA	TGAACTTCGA	GTTGCTCCAG	AAGAACGTCC	TGTTCTGCTT	ACTGAAGCAC
pt66	TGGCATCATA	CCTTTTACAA	TGAACTTCGT	GTGGCTCCAG	AGGAGCACCC	AGTTCTCCTC	ACAGAAGCTC
pt79	TGGCATCATA	CCTTTTACAA	TGAACTTCGT	GTGGCTCCAG	AGGAGCACCC	AGTTCTCCTC	ACAGAAGCTC
pt82	TGGCATCATA	CTTTTTACAA	CGAGCTTCGT	GTGGCACCAG	AGGAACACCC	TGTACTGCTC	ACAGAAGCAC
PoAc101	TGGCATCATA	CCTTTTACAA	TGAGTTGCGT	GTTTCTCCTG	ATGAGCACCC	TGTGCTACTT	ACTGAAGCAC
PoAc58	TGGCATCATA	CTTTCTATAA	CGAGCTTCGT	GTTGCACCAG	AGGAGCATCC	TGTCCTCCTA	ACTGAAGCAC
PoAc71	TGGCATCATA	CATTCTATAA	TGAACTTCGA	GTTGCTCCAG	AAGAACATCC	TGTTTTGCTT	ACTGAAGCAC
PoAc75	TGGCATCACA	CTTTCTACAA	TGAACTTCGA	GTTGCTCCAG	AAGAGCATCC	GGTTCTGCTC	ACTGAGGCAC
PoAc97	TGGCATCATA	CTTTCTACAA	TGAGCTTCGT	GTTGCCCCTG	AGGAGCATCC	TGTCCTCCTA	ACTGAAGCCC
PoAc99	TGACATCATA	CTTTCTACAA	TGAGCTTCAT	GTTGCCCCCG	AGGAGCACCC	TGTTGTGCTC	ACTGAAGCAC
pt42	TGGCATCATA	CATTTTACAA	TGAACTTCGT	GTTGCTCCAG	AGGAACACCC	TGTTCTACTC	ACCGAAGCAC
pt65	TGGCATCATA	CATTTTACAA	TGAACTTCGT	GTTGTTCCAG	AGGAACACCC	TGTTCTACTC	ACCGAAGCAC
tm32	TGGCATCATA	CATTCTACAA	CGAACTTCGA	GTGGCTCCAG	AGGAACACCC	TGTTCTTCTC	ACAGAGGCAC
tm41	TGGCATCATA	CTTTCTACAA	TGAACTTCGT	GTTGCCCCTG	AGGAGCATCC	TGTCCTCCTA	ACTGAAGCCC
tm52	TGGCATCATA	CCTTTTACAA	TGAACTTCGT	GTGGCTCCAG	AGGCGCACCC	AGTTCTCCTC	ACAGAAGCTC
tm51	TGGCATCATA	CTTTCTACAA	TGAGCTTCGT	GTTGCCCCCG	AGGAGCACCC	TGTTCTGCTT	ACTGAAGCAC
tm105	TGGCATCACA	CTTTTTACAA	TGAGCTCCGA	GTTGCTCCTG	AGGACGATTC	CATTCTTTTG	ACTGAGGCAC
tb103	TGGCATCATA	CCTTCTACAA	CGAGCTTCGT	GTTGCCCACG	AGGAGCACCC	TGTTCTTCTC	ACTGAAGCAC
tb104	TGGCATCATA	CATTTTACAA	CGAGCTTCGT	GTTGCTCCCG	AGGAGCATCC	TGTTCTTCTT	ACTGAGGCAC
tb66	TGGCATCATA	CCTTTTACAA	TGAGTTGCGT	GTTGCTCCTG	AAGAGCATCC	TTGGCTACTT	ACTGAAGCAC
tb71	TGGCATCATA	CTTTCTACAA	TGAGCTTCGT	GTTGCGCCCG	AGGAGCATCC	AGTCCTCCTT	ACTGAGGCGC
tb93	TGGCATCACA	CCTTCTACAA	TGAACTTCGT	GTGGCTCCAG	AAGAACACCC	TGTTCTCCTC	ACAGAAGCTC
tob1	TGGCATCATA	CATTTTACAA	CGAACTTCGT	GTTGCGCCAG	AGGAACACCC	TGTACTACTC	ACTGAAGCAC
tb53	TGGCATCACA	CATTTTACAA	TGAGCTCCGT	GTTGCTCCCG	AGGAGCATCC	CCTTCTTCTG	ACAGAGGCAC
tb54	TGGCATCATA	CTTTCTACAA	TGAGCTTCGA	GTTGCTCCTG	AAGAACACCC	TGTTCTCCTG	ACTGAGGCAC
soy109	TGGCATCACA	CATTTTATAA	TGAACTTCGT	GTTCTGCCTG	AGGACGACCC	AGTCGTTCTC	ACTGAGGCTC
soy115	TGGCATCACA	CCTTCTACAA	TGAACTCCGT	GTCTCGCCTG	AGGAGCACCC	AGTTCTGCTC	ACCGAAGCAC
soy118	TGGCATCATA	CCTTCTATAA	TGAGCTTCGT	GTTGCCCCAG	AAGGGCATCC	AGTTCTCTTA	GCTGAGGCCC
soy119	TGGCATCACA	CTTTCTACAA	TGAACTCCGT	GTGGCGCCAA	AGGAGCACCC	CGTTCTCCTC	ACCGAAGCAC
soy57	TGGCATCACA	CATTTTACAA	TGAGCTACGT	GTTGCCCCTG	AAGAACACCC	CGTACTTCTT	ACCGAGGCAC
soy58	TGGCATCACA	CATTTTACAA	TGAATTGCGT	GTTGCTCCTG	AGGACGACCC	AGTGCTTCTA	ACTGAGGCTC
soy69	TGGCATCATA	CCTTCTATAA	TGAGCTTCGT	GTGGCCCCAG	AAGAGCATCC	AGTTCTGTTG	ACAGAGGCTC
soy70	TGGCATCACA	CCTTCTACAA	TGAACTCCGT	GTCGCTCCTG	AGGCAAGCCC	AGTTCTGCTA	ACCGAAGCAC
soy86	TGGCGTCACA	CATTTTATAA	TGAACTTCGT	GTTGTTTCCG	AGGAGCCCCA	?GTGCTCTCA	?CTGAGGCCC
SAc1	TGGCATCATA	CCTTCTATAA	TGAGCTTCGT	GTGGCCCCAG	AAGAACATCC	AGTTCTGTTG	ACGGAGGCTC
SAc3	TGGCATCACA	CATTTTATAA	TGAACTTCGT	GTTGTTTCCG	AGGAGCCCCA	?GTGCTCTCA	?CTGAGGCCC
mz56	TGGCACCACA	CCTTCTACAA	CGAGCTCCGT	GTGGCGCCTG	AGGAGCACCC	TATACTGCTG	ACCGAGGCTC
mz63	TGGCACCACA	CCTTCTACAA	CGAGCTCCGT	GTGGCGCCTG	AGGAGCACCC	CATACTGCTG	ACCGAGGCTC
mz81	TGGCACCACA	CCTTCTATAA	CGAACTCCGT	GTTGCGCCTG	AAGAGCACCC	TGTGCTGCTG	ACCGAAGCCC
mz83	TGGCACCACA	CCTTCTATAA	TGAACTCCGT	GTGGCGCCTG	AAGAGCACCC	TGTCGTGCTG	ACGGAAGCCC
mz87	TGGCATCACA	CTTTCTACAA	TGAGCTCCGT	GTTGCCCCGG	AGGAGCACCC	TGTGTTGCTT	ACTGAGGCTC
mz89	TGGCACCACA	CCTTCTACAA	TGAGCTTCGT	GTGGCACCTG	AGGAGCACCC	TACACTGCTG	ACCGAGGCTC
mz95	TGGGCACCAC	ACCTTCTACA	TGAACTGCGT	GTTGCTCCTG	AAGAGCACCC	AGTTCTGCTG	ACAGAAGCAC
MAc1	TGGCATCACA	CCTTCTACAA	CGAGCTCCGT	GTTTCGCCTG	AAGATCACCC	TGTGCTGCTG	ACCGAGGCCC

Appendix: con't

							350
vlvx	CTCTTAATCC	CAAGGCGAAC	CGTGAGAAGA	TGACGCAGAT	CATGTTCGAG	ACCTTCAAC-	-----GTCCC
PsAc2	CACTCAACCC	AAAGGCCAAC	AGAGAAAAGA	TGACCCAAAT	CATGTTTGAG	ACCTTCAAT-	-----GTACC
PsAc1	CACTCAACCC	AAAGGCCAAC	AGAGAAAAGA	TGACCCAAAT	CATGTTTGAG	ACCTTCAAT-	-----GTACC
SoAc1	CCCTGAACCC	CAAGGCTAAC	CGTGAGAAGA	TGACCCAGAT	CATGTTCGAG	ACCTTCAAC-	-----ACCCC
AAc1	CGCTTAACCC	GAAAGCTAAT	CGTGAGAAGA	TGACTCAAAT	CATGTTTGAG	ACTTTCAAT-	-----GCCCC
RAc1	CTCTCAACCC	CAAGGCCAAT	CGTGAGAAGA	TGACCCAGAT	CATGTTTGAG	ACCTTCAAC-	-----ACCCC
RAc3	CTCTGAACCA	AAGGGCTAAC	AGAGAGAAAA	TGACCCAAAT	CATGTTTGAG	ACATTCAAT-	-----GCACC
RAc7	CTATGAACCC	CAAGGCAAAC	AGAGAGAAGA	TGACCCAGAT	CATGTTCGAG	ACCTTCAAC-	-----TGCCC
RAc2	CGCTTAACCC	TAAGGCCAAC	AGGGAGAAGA	TGACACAAAT	TATGTTTGAG	ACATTCACGT	TGCACGTTCC
DcAc1	CACTGAACCC	AAAAGCAAAT	AGGGAAAAGA	TGACCCAGAT	CATGATTGAG	ACCTTTAAT-	-----GTGCC
pt46	CTCTCAACCC	TAAGGCTAAT	CGCGAGAAAA	TGACTCAGAT	AATGTTTGAA	ACATTCAAT-	-----GCCCC
pt66	CTCTTAACCC	AAAAGCCAAT	CGAGAGAAGA	TGACTCAGAT	CATGTTTGAA	ACCTTCAAC-	-----ACCCC
pt79	CTCTTAACCC	AAAAGCCAAT	CGAGAGAAGA	TGACTCAGAT	CATGTTTGAA	ACCTTCAAC-	-----ACCCC
pt82	CTCTTAACCC	GAAGGCTAAT	CGTGAGAAAA	TGACTCAAAT	CATGTTCGAG	ACATTCAAT-	-----ACTCC
PoAc101	CCTTGAACCC	CAAGGCCAAT	AGAGAGAAAA	TGACACAGAT	AATGTTTGAG	ACTTTCAAT-	-----GTTCC
PoAc58	CTCTTAATCC	AAAAGCTAAT	CGTGAAAAAA	TGACACAGAT	TATGTTTGAG	ACTTTTAAT-	-----ACTCC
PoAc71	CTCTCAACCC	TAAGGCTAAT	CGCGAGAAAA	TGACTCAGAT	AATGTTTGAA	ACATTCAAT-	-----GCCCC
PoAc75	CTCTCAATCC	AAAGGCTAAT	CGTGAGAAAA	TGACTCAAAT	CATGTTCGAA	ACATTCAAT-	-----?????
PoAc97	CTCTTAACCC	CAAGGCTAAT	CGTGAAAAGA	TGACCCAGAT	TATGTTTGAG	ACTTTCAAT-	-----ACCCC
PoAc99	CTCTCAACCC	TAAG?CCAAC	AGAGAGAAAA	TGACCCAGAT	TATATTT?AG	ACCTTCAAT-	-----GTTCC
pt42	CTCTCAATCC	TAAGGCTAAT	CGTGAGAAGA	TGACTCAAAT	CATGTTTGAG	ACATTTAAC-	-----ACTCC
pt65	CTCTCAATCC	TAAGGCTAAT	CGTGAGAAGA	TGACTCAAAT	CATGTTTGAG	ACATTTAAC-	-----ACTCC
tm32	CTCTAAACCC	TAAGGCGAAT	CGCGAAAAGA	TGACTCAAAT	CATGTTTGAG	ACCTTCAAC-	-----GCGCC
tm41	CTCTTAACCC	AAAGGCTAAT	CGTGAAAAGA	TGACCCAGAT	TATGTTTGAG	ACTTTCAAT-	-----ACCCC
tm52	CTCTTAACCC	AAAAGCCAAT	CGAGAGAAGA	TGACTCAGAT	CATGTTTGAA	ACCTTCAAC-	-----ACCCC
tm51	CTCTCAACCC	TAAGGCCAAC	AGAGAGAAAA	TGACCCAGAT	TATGTTTGAG	ACCTTCAAC-	-----GTTCC
tm105	CACTTAACCC	TAAGGCGAAC	CGAGAGAAAA	TGACTCAAAT	CATGTTTGAG	ACATTCAAT-	-----GTGCC
tb103	CTCTCAACCC	TAAGGCCAAC	AGAGAGAAAA	TGACCCAGAT	TATGTTTGAA	ACATTCAAC-	-----GTTCC
tb104	CACTTAACCC	CAAGGCAAAC	CGAGAGAAAA	TGACCCAGAT	CATGTTTGAG	ACCTTCAAT-	-----GTGCC
tb66	CCTTGAATCC	AAAGGCCAAT	AGAGAGAAGA	TGACACAGAT	CATGTTTGAG	ACTTTCAAT-	-----GTACC
tb71	CTCTTAACCC	AAAGGCTAAT	CGTGAAAAGA	TGACCCAGAT	TATGTTTGAG	ACTTTTAAT-	-----ACCC?
tb93	CTCCTAATCC	AAAGGCCAAT	CGAGAAAAGA	TGACTCAGAT	CATGTTTGAG	ACCTTCAAC-	-----GCCCC
tob1	CTCTTAACCC	GAAGGCTAAT	CGCGAAAAGA	TGACTCAAAT	CATGTTTGAG	ACATTTAAT-	-----ACTCC
tb53	CGCTTAACCC	TAAGGCAAAC	CGAGAGAAAA	TGACTGAAAT	CATGTTCGAG	ACATTCAAT-	-----GTGCC
tb54	CACTTAATCC	TAAGGCCAAC	AGAGAGAAGA	TGACCCAGAT	TATGTTTGAG	ACATTTAGC-	-----GTTCC
soy109	CTCTCAACCC	AAAGGCCAAC	AGAGAAAAGA	TGACCCAAAT	CATGTTTGAG	ACCTTCAAT-	-----GTGCC
soy115	CTCTCAACCC	AAAGGCTAAT	CGGGAGAAAA	TGACCCAAAT	TATGTTTGAG	ACCTCCAAT-	-----ACCCC
soy118	CTCTTAACCC	TAAGGCTAAT	CGTGAGAAGA	TGACCCAAAT	CATGTTTGAG	ACCTTTAAC-	-----ACCCC
soy119	CTCTCAACCC	AAAGGCTAAT	CGTGAGAAAA	TGACCCAAAT	TATGTTTGAG	ACCTTCAAC-	-----ACACC
soy57	CACTCAATCC	CAAGGCCAAC	AGAGAGAAGA	TGACCCAAAT	TATGTTTGAG	ACTTTCAAT-	-----GTGCC
soy58	CACTCAACCC	AAAGGCCAAC	CGAGAAAAAA	TGACCCAAAT	CATGTTTGAG	ACCTTCAAT-	-----GTGCC
soy69	CTCTTAACCC	TAAGGCTAAC	CGTGAGAAGA	TGACCCAAAT	TATGTTTGAG	ACCTTTAAT-	-----GCCCC
soy70	CTCTCAACCC	AAAGGCTAAT	CGTGAGAAAA	TGACCCAAAT	TATGTTTGAG	ACCTTCAAT-	-----ACCCC
soy86	CCCTCAACCC	AAAGGCCAAC	AGAGAAAAGA	TGACCCAAAT	CATGTTTGAG	ACCTTCAAT-	-----GTGCC
SAc1	CTCTTAACCC	TAAGTATAAC	CGTGAGAAGA	TG?CCCGGGT	TATGTTTGAG	ACCTTTGGT-	-----GCCCC
SAc3	CCCTCAACCC	AAAGGTCAAC	---AGAGAAA	GTGCCCAAAT	CATGTTTGAG	ACCTTCAAT-	-----GTGCC
mz56	CTCTGAACCC	CAAGGCCAAC	AGGGAGAAGA	TGACCCAGAT	CATGTTTGAG	ACGTTTAGC-	-----TGCCC
mz63	CTCTGAACCC	CAAGGCCAAC	AGGGAGAAGA	TGACCCAGAT	CATGTTTGAG	ACGTTTAGC-	-----TGCCC
mz81	CACTCAACCC	AAAGGCTAAC	AGGGAGAAGA	TGACCCAAAT	CATGTTTGAG	ACTTTCAAT-	-----GTTCC
mz83	CACTCAACCC	CAAAGCTAAC	AGGGAGAAGA	TGACCCAAAT	CATGTTTGAG	ACCTTCAAT-	-----GTTCC
mz87	CTTTGAACCC	AAAGGCCAAC	AGGGAGAAGA	TGACCCAGAT	TATGTTCGAG	ACTTTCAAT-	-----GTTCC
mz89	CTCTGAACCC	CAAGGCAAAC	AGGGAGAAGA	TGACCCAGAT	TATGTTCGAG	ACATTCAAC-	-----TGCCC
mz95	CTCTGAACCC	AAAGGCAAAC	AGAGAGAAGA	TGACCCAAAT	TATGTTCGAG	ACATTCAAT-	-----GCACC
MAc1	CTCTCAACCC	CAAGGCCAAC	AGAGAGAAAA	TGACGCAGAT	TATGTTTGAA	ACCTTCGAA-	-----TGCCC

Appendix: con't

```
                                                                                                                    420
vlvx      TGCCATGTAT GTGGCGATCC AGGCCGTGCT GTCGCTGTAT GCTAGCGGTC GTACGACGGG TATCGTTCTA
PsAc2     TGCCATGTAT GTGGCCATCC AGGCTGTCCT CTCCCTCTAT GCAAGTGGTC GTACAACAGG TATTGTCTTG
PsAc1     TGCCATGTAT GTGGCCATCC AGGCTGTCCT CTCCCTCTAT GCAAGTGGTC GTACAACAGG TATTGTCTTG
SoAc1     CGCCATGTAC GTCGCCATCC AGGCCGTCCT CTCTCTGTAT GCCAGCGGTC GTACCACAGG TATCGTGCTC
AAc1      TGCTATGTAT GTGGCTATTC AGGCTGTTCT TTCTCTTTAT GCCAGTGGTC GTACTACCGG TATTGTGCTC
RAc1      TGCTATGTAC GTCGCCATCC AGGCCGTCCT CTCTCTGTAT GCCAGTGGTC GTACCCAAGG TATTGTGTTG
RAc3      AGCAATGTAT GTGGCTATTC AAGCTGTTCT TTCACTATAT GCCAGTGGTC GTACAACAGG TATGGTGCTT
RAc7      AGCGATGTAC GTTGCCATCC AGGCCGTTCT TTCGTTGTAT GCCAGTGGTC GAACATCTGG TATTGTACTT
RAc2      AGCCATGTAT GTCGCTATTC AAGCCGTGCT TTCCCTCTAT GCTAGTGGAC GTACTACTGG TATTGTCTTG
DcAc1     TGCCATGTAT GTTGCTATCC AGGCTGTGCT GTCTCTATAT GCTAGTGGCC GCACAACAGG TATTGTTTTG
pt46      AGCAATGTAT GTTGCTATCC AAGCTGTTTT GTCTTTGTAT GCTAGTGGAC GTACAACTGG TATTGTGCTG
pt66      TGCCATGTAT GTTGCCATTC AGGCTGTTCT TTCCCTCTAT GCCAGTGGAC GTACAACTGG AATTGTGCTG
pt79      TGCCATGTAT GTTGCCATTC AGGCTGTTCT TTCCCTCTAT GCCAGTGGAC GTACAACTGG AATTGTGCTG
pt82      CGCTATGTAT GTTGCCATTC AAGCTGTTCT ATCCCTTTA" GCCAGTGGTC GTACAACAGG TATTGTGATG
PoAc101   TGCCATGTAC GTCGCTATTC AGGCAGTCTT GTCTCTTTAC GCCAGTGGCC GTACAACAGG TATTGTGCTG
PoAc58    AGCTATGTAT GTTGCTATAC AAGCTGTTCT TTCGCTGTAT GCCAGTGGTC GTACCACCGG TATTGTGTTG
PoAc71    TGCAATGTAT GTTGCTATTC AAGCTGTTTT GTCTTTGTAT GCTAGTGGAC GTACAACTGG TATTGTGCTG
PoAc75    ?????????? ?????????? ?????????? ?????????? ?????????? ??????????G TAGTGTTTTG
PoAc97    AGCTATGTAT GTTGCTATTC AGGCTGTACT CTCACTGTAT GCCAGTGGTC GTACCACCGG TATTGTGTTG
PoAc99    GGCTATGTAT GTTGCTATTC AGGCTGTGCT TTCCTTGTAT GCTAGTGGTC GTACAACTGG TATTGTGTTG
pt42      TGCTATGTAT GTCGCCATTC AAGCTGTTCT CTCTCTTTAT GCCAGTGGTC GTACAACTGG TATTGTCATG
pt65      TGCTATGTAT GTCGCCATTC AAGCTGTTCT CTCTCTTTAT GCTAGTGGTC GTACAACTGG TATTGTCATG
tm32      TGCAATGTAT GTCGCGATTC AAGCCGTACT ATCTCTATAT GCCAGTGGCG GTACAACAGG GATTGTCATG
tm41      AGCTATGTAT GTTGCTATTC AGGCTGTACT CTCACTGTAT GCCAGTGGTC GATCCACCGG TATTGTGTTG
tm52      TGCCATGTAT GTTGCCATCC AGGCTGT?CT TTCCCTCTAT GCCAGTGGAC GTACAACTGG TATTGTGCTG
tm51      AGCTATGTAT GTTGCTATTC AGGCTGTGCT TTCCTTGTAT GCTAGTGGTC GTACAACTGG TATTGTGTTG
tm105     AGCCATGTAT ATTGCAATCC AGGCTGTTCT ATCTCTTTAT GCTAGTGGTC GTACAACTGG TATCGTGCTT
tb103     GGCTATGTAT GTTGCTATTC AGGCTGTCCT TTCCTTGTAT GCTAGTGGTC GTACAACTGG TATTGTGTTG
tb104     TGCCATGTAT GTTGCAATCC AAGCCGTTCT TTCTCTGTAT GCTAGTGGTC GTACTACTGG TATTGTGCTT
tb66      TGCCATGTAT GTTGCTATTC AAGCTGTGTT GTCCCTATAC GCCAGTGGCC GTACAACAGG TATTGTGTTG
tb71      AGCTATGTAT GTTGCTATAC AGGCTGTACT CTCACTGTAT GCCAGTGGTC GTACCAC?GG TATTGTGTTG
tb93      TGCCATGTAT GTCGCCATTC AGGCTGTTCT TTCCCTCTAT GCCAGTGGAC GTACTACTGG TATTGTGCTG
tob1      TGCTATGTAT GTCGCCATTC AAGCCGTTCT ATCCCTTTAT GCCAGTGGCC GTACAACAGG TATTGTCCTG
tb53      AGCCATGTAT GTCGCGATCC AGGCTGTTCT CTCTCTGTAT GCTAGTGGTC GTACAACTGG CATTGTGCTT
tb54      AGCTATGTAT GTTGCTATTC AGGCCGTGCT CTCATTGTAT GCTAGCGGTC GTACCACTGG TATTGTGTTG
soy109    TGCCATGTAT GTGGCCATCC AGGCTGTTCT CTCCTTGTAT GCCAGTGGTC GCACAACTGG TATTGTGTTG
soy115    TGCCATGTAT GTTGCCATCC AGGCTGTTCT ATCCTTGTAT GCCAGTGGCC GTACAACTGG TATCGTTCTA
soy118    CGCTATGTAT GTAGCCATCC AGGCTGTTCT TTCACTTTAT GCAAGTGGTC GTACAACTGG TATTGTGTTG
soy119    TGCTATGTAT GTTGCCATCC AGGCTGTTTT ATCTCTCTAT GCTAGCGGCC GTACAACTGG TATTGTTCTG
soy57     TGCCATGTAT GTTGCAATTC AGGCTGTCTT GTCTCTGTAT GCCAGTGGAC GTACTACAGG TATTGTGCTT
soy58     TGCCATGTAT GTGGCCATCC AGGCTGTCCT CTCCCTATAT GCAAGTGGTC GCACAACTGG TATTGTTTTG
soy69     TGCTATGTAT GTTGCTATTC AGGCTGTTCT TTCACTGTAT GCCAGTGGTC GTACAACTGG TATTGTGTTG
soy70     TGCAATGTAT GTTGCCATCC AGGCTGTTCT ATCCTTGTAT GCCAGTGCGC GTACAACTGG TATCGTTCTA
soy86     TGCCATGTAT GTGGCCATCC AGGCTGTTCT CTCCTTGTAT GCAAGTGGTC GCACAACTGG TATTGTATTG
SAc1      TGCTGTGTGT GTTGCTGTTC TGGCTGTTCT TTCACTGTAT GCCAGTGGCC GTACGACTGG TATTGTGTTG
SAc3      TGCCATGTAT GTGGCCATCC AAGCTGTTCT CTCCTTGTAT GCAAGTGGTC GCACAACTGG TATTGTATTG
mz56      AGCAATGTAT GTGTCCATCC AGGCAGTTCT TTCCCTGTAC GCCAGTGGAC GAACGACTGG TATCGTGCTC
mz63      AGCAATGTAT GTGTCCATCC AGGCAGTTCT TTCCCTGTAC GCCATGGGAC GAACGACTGG TATCGTGCTC
mz81      TGCCATGTAT GTCGCCATCC AGGCTGTGCT TTCCCTGTAC GCCAGTGGAC GTACAACTGG TATTGTATTG
mz83      TGCCATGTAT GTCGCCATCC AGGCTGTGCT TTCCCTGTAC GCCAGTGGGC GTAACACTGG TATTGTGTTG
mz87      TGCCATGTAC GTTGCAATTC AAGCAGTCCT GTCACTCTAT GCTAGTGGCC GTACTACTGG TATTGTTCTC
mz89      AGCAATGTAT GTGGCCATCC AGGCTGTTCT TTCCCTGTAC GCCAGTGGTC GAACAACGGG TATTGTGCTC
mz95      AGCAATGTAT GTTGCTATCC AAGCTGTCCT TTCCCTGTAT GCCAGCGGTC GCACTACAGG TATTGTGCTC
MAc1      AGCAATGTAT GTTGCTATCG AGGCTGTTCT TTCGTTGTAC GCTAGTGGGC GAACAACTGG TATTGTGATG
```

Appendix: con't

```
                                                                                                                    490
vlvx      GATTCCGGTG ACGGTGTAAC CCACACGGTG CCCATCTATG AAGGCTACGC GCTGCCGCAC GCCATCCTGC
PsAc2     GATTCTGGTG ATGGTGTGAG TCATACCGTG CCAATCTATG AGGGTTATGC ACTCCCTCAT GCCATCCTTC
PsAc1     GAGTCTGGTG ATGGTGTGAG TCATACCGTG CCAATCTATG AGGGTTATGC ACTCCCTCAT GCCATCCTTC
SoAc1     GACTCGGGAG ATGGTGTCAG CCACACTGTC CCCATCTACG AAGGGTACGC CCTCCCCCAC GCCATCCTGC
AAc1      GACTCTGGAG ATGGTGTGAG CCACACTGTT CCTATCTATG AGGGGTATGC ACTTCCACAT GCTATCCTAC
RAc1      GACTCTGGTG ATGGTGTCAG CCACACTGTC CCCATCTATG AAGGATATGC TCTCCCCCAT GCTATCCTTC
RAc3      GATTCTGGTG ATGGTGTGAC TCACACAGTG CCAATCTATG AAGGATATGC ACTTCCTCAT GCTATTCTTC
RAc7      GACTCTGGTG ATGGTGTAAG TCACACTGTT CCAATATACG AAGGATTTAC ACTCGCGCAT GCTATTCTTC
RAc2      GATTCTGGAG ATGGTGTCAG TCACACAGTC CCAATCTACG AAGGTTATGC CCTTCCGCAT GCCATTCTCC
DcAc1     GACTCTGGTG ATGGTGTTAG CCACACGGTG CCAATTTATG AAGGTTATGC ACTTCCACAT GCAATTCTTC
pt46      GATTCTGGAG ATGGTGTGAG CCACACAGTT CCTATATATG AAGGTTATGC ACTACCTCAT GCCATCCTAC
pt66      GACTCTGGAG ATGGTGTCAG CCACACAGTT CCCATCTATG AAGGTTATGC CCTCCCACAC GCCATCCTGC
pt79      GACTCTGGAG ATGGTGTCAG CCACACAGTG CCCATCTACG AAGGTTATGC CCTCCCACAC GCCATCCTGC
pt82      GATTCTGGTG ATGGCGTCAG TCACACTGTC CCAATCTATG AGGGATATGC TTTACCACAT GCTATTTTAC
PoAc101   GATTCTGGTG ATGGTGTCAG TCATACAGTG CCAATTTATG AAGGCTATGC CCTTCCTCAT GCAATTCTTC
PoAc58    GACTCTGGTG ATGGTGTCAG CCACACTGTC CCAATTTATG AGGGATATGC CCTCCCACAT GCCATTCTTC
PoAc71    GATTCTGGAG ATGGTGTAAG CCACACAGTT CCTATATATG AAGGTTATGC ACTACCTCAT GCCATCTTAC
PoAc75    GATTCCGGAG ATGGTGTTAG CCACACAGTT CCTATCTACG AAGGTTATGC ACTTCCTCAC GCGATCCTTC
PoAc97    GACTCTGGTG ATGGTGTCAG CCACACTGTC CCAATTTATG AAGGGTATGC CCTTCCACAT GCCATTCTCC
PoAc99    GACTGTGGTG ATGGGGTGAG TCACACTGTC CCTATGTATG AGGGTTATCT TTGCCGTCAT GCCATTCTTC
pt42      GACTCTGGTG ATGGTGTTAG CCACACTGTC CCAATCTATG AGGGATATGC TCTCCCACAT GCTATCCTTC
pt65      GGCTCTGGTG ATGGTGTTAG TCACACTGTC CCAATCTATG AGGGATATGC TCTCCCACAT GCTATCCTTC
tm32      GACTCTGGGG ATGGGGTGAG TCACACCGTA CCAATCTATG AGGGATATGC TCTACCACAT GCTATCCTTC
tm41      GACTCTGGTG ATGGTCTCAG CCACACTGTC CCAATTTATG AAGGGTATGC CCTTCCACAT GCCATTCTCC
tm52      GACTCTGGAG ATGGTGTCAG CCACACAGTT CCCATCTATG AAGGTTATGC CCTCCCACAC GCCATCCCGC
tm51      GACTCTGGTG ATGGTGTGAG TCACACTGTC CCTATTTACG AGGGTTATGC TTTGCCGCAT GCCATTCTTC
tm105     GATTCTGGTG ATGGTGTGAC CCACACTGTC CCTATCTATG AGGGATATGC ACTTCCCCAT GCTATCTTGC
tb103     GACTCTGGTG ATGGTGTGAG TCACACTGTC CCTATCTACG AGGGTTATGC TTTGCCTCAT GCTATTCTTC
tb104     GATTCTGGTG ATGGTGTGAG CCACACTGTC CCCATCTATG AGGGATATGC ACTTCCCCAT GCCATCTTGC
tb66      GATTCCGGCG ACGGTGTCTC ACATACAGTA CCAATTTATG AAGGTTACGC CCTTCCTCAT GCAATTCTTC
tb71      GACTCTGGTG ATGGTGTCAG CCACACTGTC CCAATTTATG AGGGGTATGC CCTCCCACAT GCCATTCTCC
tb93      GACTCTGGTG ATGGTGTCAG CCACACAGTT CCCATCTATG AAGGTTATGC TCTCCCACAC GCAATCCTGC
tob1      GATTCTGGTG ATGGTGTTAG CCACACTGTC CCAATTTATG AGGGATATGC TTTGCCACAT GCTATCCTCC
tb53      GACTCTGGTG ATGGTGTGAG CCACACTGTC CCTATCTACG AGGGATATGC ACTTCCCCAT GCTATCTTGC
tb54      GACTCTGGTG ATGGTGTGAG TCACCACGTC CCTATCTATG AAGGTTATGC TTTGCCCCAT GCCATTCTTC
soy109    GACTCTGGTG ATGGTGTGAG TCACACTGTA CCAATCTATG AGGGGTATGC ACTCCCCCAT GCCATCCTTC
soy115    GACTCTGGAG ACGGTGTGAG CCATACCGTC CCCATTTACG AGGGCTACGC TCTCCCTCAT GCCATCCTCC
soy118    GATTCTGGTG ATGGTGTCAG CCACACTGTC CCTATCTATG AGGGTTATGC CCTCCCACAT GCCATCCTTC
soy119    GACTCTGGAG ATGGTGTCAG CCACACTGTG CCCATTTATG AAGGGTATGC CCTCCCTCAT GCCATCCTCC
soy57     GATTCTGGTG ATGGTGTGAG CCACACGGTG CCCATATATG AAGGGTATGC TCTTCCTCAT GCAATTCTTC
soy58     GACTCTGGTG ATGGTGTGAG TCACACTGTG CCAATCTATG AGGGCTATGC ACTCCCTCAT GCCATTCTCC
soy69     GATTCTGGGG ATGGTGTCAG CCACACAGTC CCTATCTATG AGGGTTATGC CCTCCCACAT GCAATCCTTC
soy70     GACTCTGGAG ATGGTGTCAG CCATACTGTC CCCATTTACG AGGGGTATGC TCTCCCTCAT GCCATCCTCC
soy86     GACTCTGGTG ATGGTGTGAG TCACACTGTA CCAATCTATG AGGGGTATGC ACTCCCCCAT GCCATCCTTC
SAc1      GATTCTGGGG ATGGTGTCAG CCACACAGTG CCTATCTATG AGGGTTATGC CCTCCCACAT GCAATCCTTC
SAc3      GACTCTGGTG ATGGTGTGAG TCACACTGTA CCAATCTATG AGGGGGATGC ACTCCCCCAT GCCATCCTTC
mz56      GACTCTGGTG ACGGCGTGCA GCACTGTCCC CCAATCTACG AAGGGTACAC GCTTCCTCAT GCTATTATTC
mz63      GACTCTGGTG ACGGCGTGCA GCACTGTCCC CCAATCTACG AGGGGTACAC GCTTCCTCAT GCTATTATTC
mz81      GACTCTGGTG ATGGTGTCAG CCACACAGTG CCAATCTATG AAGGTTATGC CCTTCCTCAT GCCATCCTGC
mz83      GACTCTGGTG ATGGTGTCAG CCACACCGTG CCAATCTACG AAGGTTATGC CCTTCCTCAT GCCATCCTGC
mz87      GATTCTGGTG ATGGTGTGAG CCATACTGTG CCCATTTATG AAGGATATGC CCTTCCTCAT GCTATTCTAC
mz89      GACTCTGGTG ATGGTGTGAG CCACACCGTC CCTATCTACG AAGGGTACAC GCTTCCTCAT GCTATTCTTC
mz95      GATTCTGGTG ATGGTGTGAC CCACACCGTG CCAATCTACG AGGGATACGC GCTCCCTCAT GCCATTCTTA
MAc1      GATTCAGGTG ATGGTGTGAG CCACACGGTT CCAATTTATG AAGGATACAC ACTTCCTCAT GCTATTCTCC
```

Appendix: con't

```
                                                                                                                  560
vlvx       GTCTGGACCT TGCCGGCCGT GACCTGACGG ACTACCTGAT GAAGATTCTA ATGGAGCGTG GTTACTCATT
PsAc2      GTTTGGATCT TGCTGGTCGT GATCTAACTG AATCTTTGAT GAAGATCCTC ACTGAGAGAG GGTATATGTT
PsAc1      GTTTGGATCT TGCTGGTCGT GATCTAACTG AATCTTTGAT GAAGATCCTC ACTGAGAGAG GGTATATGTT
SoAc1      GTCTCGACCT CGCTGGCCGC GACCTTACCG ACTACCTCAT GAAGATCCTG ACTGAGCGCG GCTACTCCTT
AAc1       GTCTTGATCT TGCTGGTCGT GACCTCACGG ATGCGCTGAT GAAGATCCTA ACCGAGCGTG GTTACTCATT
RAc1       GTCTCGACCT TGCTGGGCGT GATCTCACTG ATTACCTCAT GAAGATCCTG ACGGAGCGTG GTTACTCATT
RAc3       GGTTGGATCT TGCTGGTCGT GATCTGACTG ACTGCCTAAT GAAGATCCTT ACAGAAAGAG GTTATTCCTT
RAc7       GACTCGATCT TGCTGGGCGT GACCTTACCG ACAACCTTAT GAAGATTCTC ACAGAGAGGG GTTACTCCTT
RAc2       GTCTTGATCG TGCTGGTAGG GATCTGACTG ATTCCCTCAT GAAGATCCTG ACTGAGAGGG GTTACTCAGT
DcAc1      GGTTGGACCT TGCTGGCCGT GATCTTACAG ATGGCTTGAT GAAGATTCTT ACAGAAAAGA GTATATGTCA
pt46       GTCTCGACCT TGCTGGTGCT GACCTTACTG AGTACATGGT GAAGATTCTT ACTGAACGCG GATACTCCTT
pt66       GTCTTGATCT TGCAGGGCGT GATCTCACAG ATCACCTCAT GAAGATCCTC ACAGAACGAG GCTACTCTTT
pt79       GTCTTGATCT TGCAGGGCGT GATCTCACAG ATCACCTCAT GAAGATCCTC ACAGAACGAG GCTACTCTTT
pt82       GTCTTGATTT GGCTGGTCGT GACCTTACTG ATCACTTAAT GAAAATCTTG ACGGAGCGTG GTTACTCATT
PoAc101    GATTGGACCT AGCTGGGCGT GATCTTACTG ATTGTCTGAT GAAAATCCTG ACCGAAAGAG GTTATTCATT
PoAc58     GTCTCGACTT GGCAGGTCGT GACCTCACTG ATCATTTGAT GAAGATCCTC ACTGAGCGTG GTTACTCATT
PoAc71     GTCTCGACCT TGCTGGTCGT GACCTTACTG AGTACATGGT GAAGATTCTT ACTGAACGCG GATACTCTTT
PoAc75     GTCTTGATCT TGCTGGTCGT GACCTCACAG AGCATTTGGC AAAGATCCTC ACAGAGCGTG GATACTCATT
PoAc97     GTCTTGACTT GGCGGGTCGT GACCTCACTG ATAGTTTGAT GAAGATCCTC ACCGAGCGTG GTTACTCGTT
PoAc99     GTTTGGATCT TTCTGGCCGT GATCCAACTG ATAACCTGAT GAAGATTCTC ACTGAGAGAG GTTACATGTT
pt42       GTCTCGATTT AGCAGGACGT GACCTAACTG ATCACCTAAT GAAGATCTTG ACAGAGCGCG GTTACTCATT
pt65       GTCTCGATTT AGCAGGACGT GACCTAACTG ATCACCTAAT GAAGATCTTG ACAGAGCGTG GGTACTCGTT
tm32       GTCTCGATTT AGCTGGACGG GACCTGACGG ATCACCTGAT GAAGATCTTG ACAGAGCGTG GTTACTCGTT
tm41       GTCTTGACTT GGCAGGACGT GACCTCACTG ATAGTTTGAT GAAGATCCTG ACCGAGCGTG GTTACTCGTT
tm52       GTCTCGATCT TGCAGGGCGT GATCTCACAG ATCACCTCAT GAAGATCCTC ACCGAGAGAG GTTACATGTT
tm51       GTTTGGATCT TGCTGGTCGT GATTTAACTG ATAACCTGAT GAAGATCCTC ACCGAGAGAG GTTACATGTT
tm105      GGTTAGATCT TGCTGGTCGT GATCTTACAG ATTACCTCAT GATGATTCTC ACAGAGAGAG GTTATTCATT
tb103      GTTTGGATCT TGCTGGCCGT GACCTAACTG ATAACCTGAT GAAGATCCTC ACTGAGAGAG GTTACATGTT
tb104      GGTTAGATCT TGCCGGTCGT GGTCTTACTG ATTATCTCAT GAAGATTCTC ACTGAGAGAG GCTATATGTT
tb66       GGTTGGATCT TGCTGGTCGT GATCTTACTG ATTGTTTGAT GAAAATCTTG ACGGAAAGAG GTTATTCATT
tb71       GTCTTGACTT GGCAGGTCGT GACCTCACTG ATAGTTTGAT GAAGATCCTT ACCGAGCGTG GTTACATGTT
tb93       GTCTTGATCT GGCAGGGCGT GATCTCACAG ATCACCTCAT GAAGATTCTC ACAGAAAGAG GCTACTCTTT
tob1       GTCTTGATTT AGCTGGTCGT GACTTGACTG ATCATCTAAT GAAAATCTTA ACAGAGCGAG GCTACTCGTT
tb53       GGTTAGATCT TGCTGGTCGT GATCTTACAG ATTATCTCAT GAAGATTCTC ACAGAGAGAG GTTATTCATT
tb54       GTTTGGATCT TGCTGGTCGC GACCTTACTG ATAGCCTGAT GAAGATCCTC ACTGAGAGAG GATACATGTT
soy109     GTTTGGATCT CGCAGGTCGT GATCTAACTG ATCATTTGAT GAAGATCCTC ACTGAGAGAG GGTACATGTT
soy115     GTCTTGACTT AGCAGGACGT GATCTCACTG ATGCCTTGAT GAAAATCCTT ACAGAACGTG GTTATTCCTT
soy118     GTCTTGATCT GGCTGGTCGT GACGTCACAG ATTTCTTGAT GAAAATCTTG ACTGAACGTG GTTATTCCTT
soy119     GTCTTGACTT AGCAGGGCGT GACCTCACTG ATGCCCTGAT GAAAATCCTT ACTGAGCGTG GCTATTCCTT
soy57      GGTTGGACCT TGCTGGCCGT GATTTAACAG AGTACTTGGT GAAAATCCTT ACCGAGAGAG GATACTCCTT
soy58      GTTTGGATCT TGCTGGCCGC GATCTGACTG ACTCTTTGAT GAAGATCCTC ACCGAGAGAG GTTACATGTT
soy69      GTCTTGATTT GGCTGGCCGT GACCTCACAG ATTTCTTGAT GAAGATTTTG ACTGAACGTG GATATTCTTT
soy70      GTCTTGACTT AGCAGGACGT GATCTCACTG ATGCCTTGAT GAAAATCCTT ACAGAACGTG GTTATTCCTT
soy86      GTTTGGATCT CGCTGGTCGT GATCTAACTG ATCATTTGAT GAAGATCCTC ACTGAGAGAG GATACATGTT
SAc1       GTCTTGATTT GGCTGGTCGT GACCTCACAG ATTTCTTGAT GAAGATTCTG ACTCAACGTG GATATTCTTT
SAc3       GTTTGGATCT CGCTGGTCGT GATCTAACTG ATCATTTGAT GAAGATCCTC ACTGAGAGAG GATACATGTT
mz56       GATTGGACCT TGCCGGTCGT GACCTCACCG ACAACCTGAT GAAGATCCTC ACCGAGAGGG GTTACTCCTT
mz63       GATTGGACCT TGCCGGTCGT GACCTCACCG ACAACCTGAT GAAGATCCTC ACCGAGAGGG GTTACTCCTT
mz81       GCCTTGACCT TGCTGGGCGT GACCTGACTG ACAGCCTGAT GAAGATCCTC ACTGAGAGAG GTTACTCCTT
mz83       GTCTTGACCT TGCTGGGCGT GACCTGACTG ACAGCCTGAT GAAGATCCTC ACCGAGAGAG GTTACTCCTT
mz87       GTCTGGATCT TGCTGGTCGG GACCTCACCG ACTCGCTCAC GAAAATTCTC ACTGAGAGGG GCTATTCCTT
mz89       GATTGGACCT TGCTGGTCGT GACCTTACCG ACAACCTGAT GAAGATCCTT ACTGAGAGGG GTTACTCCTT
mz95       GGCTGGATCT TGCTGGTCGT GATCTGACTG ACTGCCTAAT GAAGATCCTC ACGGAAAGGG GCTATTCCTT
MAc1       GTTTGGATCT TGCGGGGCGT GACCTCACCG ACCACCTAAT GAAGATCCTC ACAGAGAGAG GGTACTCCCT
```

Appendix: con't

630

```
vlvx      CACGACCACT GCTGAGCGTG AAATCGTGCG CGACATCAAG GAGAAGCTCT GCTATGTGGC GCTGGATTTC
PsAc2     CACCACCTCG GCTGAGCGGG AAATTGTCCG TGACATAAAA GAGAAGCTTG CCTACGTTGC TCTGGATTAT
PsAc1     CACCACCTCG GCTGAGCGGG AAATTGTCCG TGACATAAAA GAGAAGCTTG CCTACGTTGC TCTGGATTAT
SoAc1     CACCACCACT GCTGAGCGGG AAATTGTCAG GGACATGAAG GAGAAGCTCG CCTACATTGC CCTGGACTAC
AAc1      CACCACCACA GCAGAGCGTG AAATTGTCAG AGACATAAAG GAGAAGCTTT GCTACATTGC TCTTGACTAC
RAc1      CACCACAACG GCCGAGCGGG AAATTGTGAG GGACATGAAG GAGAAGCTTT CCTACATCGC CCTGGACTAT
RAc3      TACAACAACT GCAGAGCGGG AAAGTGTGAG AGACATAAAA GAGAAGCTTG CATATATTGC TCTTGATTAT
RAc7      CACCACAACT GCTGAGCGGG AAATTGTTAG AGACATAAAG GAGAAGCTCG CCTATGTTGC TCTTGATTAT
RAc2      CACCACTCCT GCCGAGCGGG AAATTGCCGC TGACATCAAG GAAAAGCTTG CATACGTCCG TCTTGACTAC
DcAc1     CTACACCACT GCCGAGCGGG AAATTGTTCG TGATATGAAG GAAAAACTTG CTTATGTTGC TCTTGATTAT
pt46      CACTACTACT GCTGAGAAGG AAATTGTAAG GGATATGAAG GAGAAATTGG CATACCTTGC TCTTGATTTC
pt66      CACCACCACA GCCGAGCGGG AAATTGTTAG GGATGTGAAG GAGAAGTTGT CTTACATTGC TCTTGACTAT
pt79      CACCACCACA GCCGAGCGGG AAATCGTTAG GGATGTGAAG GAGAAGTTGT CTTACATCGC TCTTGACTAT
pt82      TACCACTACA GCAGAAAGGG AAATTGTGAG GGATGTTAAG GAAAAACTCT CATACATTGC TCTTGATTAT
PoAc101   CACCACTAGT GCTGAACGGG AAATTGTCCG CGATATGAAA GAGAAACTAG CATATGTTGC ACTTGACTAT
PoAc58    CACCACCACA GCTGAGCGAG AAATCGTCAG GGATGTGAAA GAAAAACTCT CTTACATAGC CTTAGACTTT
PoAc71    CACTACTACT GCTGAGAAGG AAATTGTGAG GGATGTGAAG GAGAAATTGG CATACCTTGC TCTTGATTTT
PoAc75    CACCACCAGT GCTGAGAAGG AAATTGTGAG AGACATGAAG GAGAAGCTGG CATACATTGC TCTAGACTTT
PoAc97    CACCACCTCA GCTGAGAGAG AAATTGTCAG GGACGTGAAA GAAAAGCTCG CTTACATAGC TCTTGACTAT
PoAc99    CACCACCACT GCTGAGCGGT AAATTGTCTG TGACATGAAG GTAAAACTTG CCTATGTGGC TCTTGACTAT
pt42      CACCACCACT GCTGAACGAG AAATCGTGAG GGATGTTAAG GAAAAGCTTT CCTACATTGC TCTTGACTAT
pt65      CACCACCACT GCTGAACGAG AAATCGTCAG GGATGTTAAG GAAAAGCTCT CCTACATTGC TCCTGACTAT
tm32      CACAACAAGT GCTGAGAGAG AAATCGTGGG AA?TGTGAAG GAAAAGCTAT CTTACATAGC TCTCGAC?AA
tm41      CACCACCTCA GCTGAGCGAG AAATTGTCAG GGACGTGAAA GAAAAGCTCG CTTACATAGC TCTTGACTAT
tm52      CACCACCACA GCCGAGCGGG AAATTGTTAG GGATGTGAAG GAGAAGTTGT CTTACATTGC TCTTGACTAT
tm51      CACCACCACT GCTGAACGGG AAATTGTCCG CGACATGAAG GAAAAGCTTG CCTATGTGGC TCTTGACTAT
tm105     CACCACTGCT GCTGAACGGG AGATTGTTCG TGATGTGAAG GAGAAGCTTG CGTATGTCGC ACTGGATTAT
tb103     CACCACCACT GCAGAACGGG AAATTGTCCG TGACATGAAG GAGGAGCTTG CTTACGTGGC TCTTGACTAC
tb104     TACCACATCT GCTGAACGGG AAATTGTTCG TGATGTGAAG GAGAAGCTTG CCTATGTTGC ATTGGATTTT
tb66      CACCACTAGT GCTGAACGGG AAATTGTCCG CGATGTTAAA GAGAAACTGG CATATGTTGC ACTTGACTAT
tb71      CACCACCTCA GCCGAGCGTG AAATTGTCAG GGACGTGAAG GAAAAGCTCG CTTACATAGC TCTTGACTAT
tb93      CACCACCACG GCCGAGCGGG AAATTGTTAG GGATGTGAAG GAGAAGTTGG CTTACATTGC TCTTGACTAT
tob1      CACCACTACT GCTGAACGAG AAATCGTGAG GGATGTGAAG GAAAAACTCT CCTACATTGC TCTTGACTTT
tb53      TACCACTGCT GCTGAACGTG AGATTGTTCG TGATTTGAAG GAGAAGCTTG CATATGTTTC TCTGGTTTTT
tb54      CACCACCACT GCTGAACGGG AAATTGTCCG TGATATGAAA GAAAAGCTTG CCTATGTGGC TCTTGACTAT
soy109    CACCACCTCT GCTGAGCGAG AAATTGTCCG TGACATGATG GAAAAACTTG CATATGTTGC CCTAGATTAT
soy115    CACTACCACA GCTGAGCGTG AAATTGTAAG GGACGTAAAG GAAAAGCTGG CTTATATTGC CCTTGATTAT
soy118    CACCACCTCA GCCGAGCGTG AAATTGTAAG GGATGTGAAA GAAAAGCTTT CTTACATTGC CCTTGACTAC
soy119    CACTACCACA GCCGAGAGAG AAATTGTAAG GGACATGAAG GAAAAGCTGG CATATCTTGC CCTTGATTAT
soy57     CAGCACCTCT GCCGAAAAGG AAATTGTCCG TGATGTGAAG GAGAAATTGA CATACGTGGC CCTTGATTTC
soy58     CACCACCTCT GCTGAGCGGG AAATTGTTCG TGACATGAAG GAGAAGCTTG CTTATGTTGC CTTGGATTAT
soy69     TACCACCTCA GCAGAGCGTG AAATTGTGAG GGATATGAAG GAAAAGCTGG CTTACATAGC CCTCGACTAT
soy70     CACTACCACA GCTGAGCGTG AAATTGTAAG GGATGTAAAG GAAAAGCTGG CTTATATTGC CCTTGATTAC
soy86     CACCACCTCT GCTGAGCGAG AAATTGTCCG TGACATGAAG GAAAAACTTG CATATGTTGC CCTAGATTAT
SAc1      TACCACCTCA GCACAGCGTG GAATTGTGAG GGATATGAAG GAAAAGCTGG CTTACATAGC CCTCGACTGT
SAc3      CACCACCTCT GCTGAGCGAG AAATTGTCCG TGACATGAAG GAAAAACTTG CATATGTTGC CCTAGATTAT
mz56      CACCACGACC GCCGAGCGAG AGATAGTCAG GGACATCAAG GAAAAACTTG CCTACATTGC TCTTGATTAC
mz63      CACCACGACC GCCGAGCGAG AGATAGTCAG GGACATCAAG GAAAAACTTG CCTACATTGC TCTTGATTAC
mz81      CACCACCTCT GCTGAACGCG AAATTGTAAG AGACATCAAG GAAAAGCTTG CATATGTGGC TCTTGAATAC
mz83      CACCACCACG GCCGAACGGG AAATCGTACG GGACATCAAG GAAAAGCTTG CATACGTGGC TCTTGACTAT
mz87      CACAACCTCC GCCGAGCGAG AAATTGTAAG GGACATCAAG GAGAAGCTTG CGTATATAGC TCTTGACTAC
mz89      CACTACGACT GCCGAGCGAG AAATTGTCAG GGACATCAAG GAAAAACTTG CCTACGTTGC CCTTGATTAT
mz95      CACCACGACG GCAGAACGGG AAATCGTGAG GGACATAAAG GAGAAGCTGG CCTTTGTCGC CCTAGACTAT
MAc1      CACTACGAGC GCTGAGCGGG AGATTGTCAG GGACATAAAG GAGAAGCTTG CCTACGTTGC CCTTGATTAT
```

Appendix: con't

```
                                                                                                                    700
vlvx      GAGCAGGAAA TGGCAACAGC AGCCTCCAGC TCG---GCGC TTGAGAAGAC GTACGAGCTG CCTGATGGCC
PsAc2     GAACAAGAGC TTGAAACTGC AAAGAGCAGT TCC---TCTA TTGAGAAAAA CTATGAGCTT CCTGATGGAC
PsAc1     GAACAAGAGC TTGAAACTGC AAAGAGCAGT TCC---TCTA TTGAGAAAAA CTATGAGCTT CCTGATGGAC
SoAc1     GACCAGGAGA TGGAGACTGC CAAGACCAGC TCT---TCCG TGGAGAAGAG CTACGAGCTT CCTGATGGAC
AAc1      GAGCAGGAAC TCGAGACAGC CAAAACCAGC TCA---TCTG TTGAGAAGAA CTACGAGCTA CCTGATGGGC
RAc1      GACCAGGAAA TGGAGACTGC CAAGACCAGC TCC---TCCG TGGAGAAGAG CTACGAGCTT CCTGATGGAC
RAc3      GAGCAGGAGC TGAGAGTCGC CAAGAGCAGC TCC---TCTG TTGAGAAGAG CTATGAGGTT GCAGATGGAC
RAc7      GAACAGGAGC TTGATACTGC CAGGAGTAGC TCC---TCTA TTGAGAAGAG CTATGAGCTT CTGGACGGCC
RAc2      GAGCAGGAGC TTGAGACTGC AAAGAGCAGC TCA---TCAG TTGAGAAGAG CTATGAGCTG CCCGATGGAC
DcAc1     GAGCAAGAGT TGGAGACTGC CAAGAGACGC TCT---GCTG TCGAGAAAAA CTATGAACTG CCTGATGGGC
pt46      GAGCAGGAAC TGGAGACAAC CACGACAGGC TCT---GCTG TTGAGAAGAA CTATGAACTG CCTGATGGAC
pt66      GAACAAGAAA TTGAAACAGC TAAGACCAGC TCA---TCTG TTGAGAAGAG CTATGAGCTT CCTGATGGAC
pt79      GAACAAGAAA TTGAAACAGC AAAGACCAGC TCA---TCTG TTGAGAAGAG CTATGAGCTT CCTGATGGAC
pt82      GAGCAAGAAT TGGATACGTC AAAGACTAGT TCT---TCTG TTGAGAAGAG TTATGAGTTG CCTGATGGTC
PoAc101   GAGCAAGAAT TGGAGACCGC AAAGAGTAGC TCT---GCTG TTGAAAAAAG CTATGAGTTA CCAGATGGTC
PoAc58    GAACAAGAAC TCGAGACTTC AAAGACCAGC TCT---TCTG TTGAGAAGAG CTATGAGCTT CCCGATGGTC
PoAc71    GAGCAGGAAC TTGAGACAAC CAAGACAGGG TCT---GCTG TTGAGAAGAA CTATGAACTG CCTGATGGAC
PoAc75    GAGCAGGAGT TGGATATGGT TAAGTCAAGC AGC---TCTG TTGAGAAAAA CTTTGAGCTT CCAGATGGCC
PoAc97    GAACAGGAAC TTGAGACTTC AAAGACCAGC TCT---TCTG TTGAGAAGAG CTATGAGCTC CCAGATGGGC
PoAc99    GAGCAGGAGA TG?AAACTGC CAGGAGCAGC TCC---TCCA TCAAAAAGAA CTATGAATTG CCTGATGGAC
pt42      GAACAAGAGA TGATAGACAG T--------- ------ACTG TTGAGAAGAC CTACGAGTTG CCTGATGGTC
pt65      GAACAAGAGA TCATAGACAC CACTTCCTCC TCAAAAACTG TTGAGAAGAC CTATGAGTTG CCTGATGGTC
tm32      GAACAAGAGA TTGAAACCTC AAAAACTAGC TCA---AACG TGGAGAAGAG CTACGAGTTG CCTGATGGTC
tm41      GAACAGGAAC TCGAGACCTC AAAGACCAGC TCT---TCTG TTGAGAAGAG CTATGAGCTC CCAGATGGGC
tm52      GAACAGGAAA TTGAAACAGC AAAGACCAGC TCA---TCTG TTGAGAAGAG CTATGAGCTT CCTGATGGAC
tm51      GAGCAGGAGA TTGAAACTGC CAGGAGCAGC TCC---TCCA TTGAAAAGAA CTATGAATTA CCTGATGGAC
tm105     GAACAGGAGC TTGAGACTAC AAAGAJCGGT AAG---GATG CTGAGAAGAG TTACGAACTT CCAGATGGGC
tb103     GAGCAGGAGC TTGACACTGC CAAGAGCAGC TCC---TCCG TTGAGAAGAA CTATGAACTG CCAGATGGAC
tb104     GAACAGGAGC TTGAGACAGC AAAGAGTAGC TCA---TCAG TTGAGAAGAG CTATGAGCTT CCTGATGGAC
tb66      GAGCAAGAGT TGGAGACTGC AAAGAGTAGC TCT---TCTG TTGAAAAAAG CTATGAGTTA CCAGATGGGC
tb71      GAACAGGAAC TCGAGACTGC AAAGACCAGC TCT---TCTG TAGAGAAGAA CTATGAGCTC CCAGATGGGC
tb93      GAACAAGAAC TTGAAACAGC AAAGACCAGC TCA---TCCA TGGAGAAGAG CTACGAGCTG CCTGATGGAC
tob1      GAGCAGGAAA TGGAAACGTC GAAAACTAGC TCT---TCTG TTGAGAAGAG CTACGAATTG CCCGATGGTC
tb53      GAGCAGGAGC TTGAGACTTC AAAGAATAGC AAG---GATG TCGAGAAGAG CTATGAGCTT CCAGATGGGC
tb54      GAGCAGGAAC TTGAAACTGC CAGGAGCAGT TCC---TCCA TTGAAAAGAA CTATGAATTG CCAGATGGAC
soy109    GAGCAAGAAC TCGAGACTGC AAAAAGCAGT TCA---TCAG TTGAGAAAAG CTATGAGCTT CCTGATGGAC
soy115    GAGCAAGAGC TTGAAACATC AAAAACCAGC TCT---GCGT TTGAGAAGAG CTATGAGTTG CCGGATGGAC
soy118    GAGCAAGAAC TGGAGACAGC CAGGACCAGC TCG---TCTG TGGAGAAGAG CTATGAGTTG CCTGATGGGC
soy119    GAGCAAGAAC TTGAAACTTC AAAAACTAGT TCT---GCTG TTGAGAAGAG CTATGAGCTT CCTGATGGAC
soy57     GAACAAGAAA TGGAGACTAC TAAGAGTAGT TCT---GCAG TGGAAAAGAG CTATGAATTG CCTGATGGAC
soy58     GAGCAAGAAC TCGAGACTGC AAAGAGCAGT TCT---TCAG TTGAGAAAAA CTACGAGCTT CCTGATGGGC
soy69     GAGCAAGAGC TAGAGACATC CAAGACCAGC TCT---TCAG TGGAGAAGAG CTATGAGTTG CCTGATGGGC
soy70     GAGCAAGAGC TTGAAACATC AAAAACTAGC TCT---GCTG TTGAGAAGAG CTATGAGTTG CCGGATGGAC
soy86     GAGCAAGAAC TCGAGACTGC AAAAAGCAGT TCA---TCAG TTGAGAAAAG CTATGAGCTT CCTGATGGAC
pSAc1     GAGCAAGAGC TGGAGACATC CGAGACCAGC TCT---TCAG TGGAGAAGAG CTATGAGTTG CCTGATGGGC
pSAc3     GAGCAAGAAC TCGAGACTGC AAAAAGCAGT TCA---TCAG TTGAGAAAAG CCATGAGCTT CCTGATGGAC
mz56      GAGCAGGAGC TGGAGACGGC CAGGACCAGC TCC---ACCG TCGAGAAGAG CTACGAGCTG CCCGATGGCC
mz63      GAGCAGGAGC TGGAGACGGC CAGGACCAGC TCC---ACCG TCGAGAAGAG CTACGAGCTG CCCGATGGCC
mz81      GACCAGGAGC TCGAGAATGC CAAGAGCAGC TCA---TCTG TGGAGAAGAG CTACGAGCTG CCTGATGGTC
mz83      GATCAGGAGC TGGAGAACGC CAAGAGCAGC TCA---TCTG TGGAGAAGAG CTACGAGCTG CCTGATGGCC
mz87      GAGCAAGAGT TGGAGACTGC CAAGAACAGC TCC---TCAG TTGAAAAGAG CTATGAGCTA CCTGATGGCC
mz89      GAACAGGAGC TGGAGACTGC CAGGACCAGC TCT---AGTG TTGAGAAGAG CTACGAGCTG CCCGATGGCC
mz95      GAGCAGGAGC TGGAGACCGC TAAGAGAAGC TCC---TCCG TCGAGAAGAG CTACGAGCTG CCCGATGCGC
MAc1      GAACAGGAGC TGGAGACTGC CAAGAGCAGC TCC---TCTG TCGAGAAGAG CTACGAGATG CCTGATGGTC
```

Appendix: con't

```
                                                                                                    770
vlvx      AGCCAATCAC AATTGGCAAC GAGCGCTTCC GTTGCCCTGA GGTGCTGTAC AACCCTAGCC TAATCGGCAT
PsAC2     AAGTTATCAC AATCGGAGCT GAGAGGTTCC GTTGCCCAGA AGTCCTTTTC CAGCCATCTA TGATTGGAAT
PsAc1     AAGTTATCAC AATCGGAGCT GAGAGGTTCC GTTGCCCAGA AGTCCTTTTC CAGCCATCTA TGATTGGAAT
SoAc1     AGGTCATCAC CATTGCGGCG GACCGATTCC GCTGCCCTGA GGTCCTCTTC CAGCCATCCT TCATTGGGAT
AAc1      AAGTGATCAC CATTGGATCA GAGCGATTCC GTTGTCCTGA GGTTCTTTAC CAGCCATCTA TGATTGGTAT
RAc1      AGGTTATCAC CATTGGTGCT GAGCGTTTCC GCTGCCCTGA GGTCCTCTTC CAGCCTTCCT TCATAGGAAT
RAc3      AAGTCATCAC AATTGGCGCG AGAGAGGTTC AGTGCCCAGA ACTG---TTC CAGCCCTCGC TGATTGGGAT
RAc7      AGGTCATCAC CATCGGAGCA GAAAGGTTCA GGTGCCCGGA GGTGCTCTTC CAGCCATCTT TTATTGGTAT
RAc2      AGGTGATCAC CATTGGCGCG GAGCGCTTCA GATGCCCAGA GGTCATGTTC CAGCCTTCTC TCATCGGCAT
DcAc1     AAGTTATCAC CATTGGAGCT GAGAGGTTCC GCTGTCCACA AGTCCTCTTC CAGCCATCTA TGATTGGGAT
pt46      AAGTGATTAC TATTGGTGCT GAGCGTTTTC GATGTCCTGA AGTTCTTTAT CAACCTTCTT TGATCGGAAT
pt66      AGGTTATCAC CATTGGTGCT GAGCGATTCC GTTGTCCAGA AGTGCTGTTC CAACCATCAA TGATCGGAAT
pt79      AGGTTATACC CATTGGTGCT GAGCGATTCC GTTGTCCAGA AGTGCTGTTC CAACCATCAA TGATCGGAAT
pt82      AGGTTATTAC AATTGGTGCT GAGCGTTTCC GTTGCCCAGA AGTACTATTC CAGCCTTCAA TGATTGGAAT
PoAc101   AGGTGATTAC CATTGGAGCA GAGAGGTTCA GATGCCCGGA AGTCTTATTC CAGCCATCAC TTGTGGGAAT
PoAc58    AAGTCATTAC CATCGGTGCT GAACGATTCC GTTGCCCTGA GGTCCTATTC CAACCTTCAC TGATCGGAAT
PoAc71    AAGTGATTAC TATTGGTGCT GAGCGTTTTC GATGTCCTGA AGTTCTTTAT CAACCTTCTT TGATCGGAAT
PoAc75    AAGTGATTAC CATTGGAGCT GAACGTTTCC GTTGCCCAGA AGTGCTGTTC CAGCCTTCCA CAATCGGAAT
PoAc97    AGGTGATCAC CATTGGTGCT GAGCGTTTCC GGTGTCCTGA GGTCCTTTTC CAACCTTCAA TGATTGGAAT
PoAc99    AAGATATTAT CATTGGTGCT GAGAGGTTCC GTTGCCCAGA GGTCCTCTTC CAGCCATCTA TGATTGGTAT
pt42      AAGTTATCAC TATTGGTGCT GAACGTTTCC GATGCCCTGA AGTGCTTTTC CAGCCTTCAA TTATCGGAAT
pt65      AAGTCATCAC TATTGGTGCT GAACGTTTCC GATGCCCTGA AGTGCTTTTC CAACCTTCAA TTATCGGAAT
tm32      AAGTGATTAC GATTGGTGCT GAACGTTTCA GATGCCCTGA AGTGCTGTTC CAGCCTTCCA TGATCGGAAT
tm41      AGGTGATACC CATTGGGTCT GAGCGTTTCC GGTGTCCTGA GGTCCTTTTC CAACCTTCAA TGATTGGAAT
tm52      AGGTTATCAC CATTGGGTCT GAGCGATTCC GCTGTCCAGA AGTGCTGTTC CAACCATCAA TGATCGGAAT
tm51      AAGTTATTAC CATTGGTGCT GAGAGGTTCG GTTGCCCAGA GGTCCTCTTC CAGCCATCCA TGATTGGTAT
tm105     AAATAGTCAC TATTGGTGCT GAGAGGTTCC GTTGCCCTGA AGTCCTATTC CAGCCATCAT TGATTGGAAT
tb103     AAGTTATTAC CATTGGTGCC GAGAGGTTCC GGTGCCCAGA AGTCCTCTTC CAGCCATCCA TGATCGGAAT
tb104     AAGTCATCAC AATCGGTGCT GAGAGGTTCC GTTGCCCTGA GGTTCTCTTC CAGCCATCAT TTATTGGAAT
tb66      AGGTCATTAC CATTGGCTCA GAGAGGTTCA GATGTCCGGA AGTTTTATTC CAGCCATCAC TTGTTGGAAT
tb71      AGGTGATCAC CATTGGAGCT GAGCGTTTCC GTTGTCCTGA GGTCCTTTTC CAACCATCAA TGATTGGAAT
tb93      AGGTTATCAC CATTGGTGCT GAGAGATTCC GCTGCCCAGA AGTTCTGTTC CAACCATCAA TGATCGGAAT
tob1      AAGTGATTAC CATCGGTGCT GAACGTTTCC GATGCCCTGA AGTCCTTTTC CAACCTTCAA TGATTGGAAT
tb53      AAATAATCAC AATTGGCGCG GAGAGGTTCC GCTGTCCTGA GGTCCTCTTC CAGCCATCAT TTATCGGAAT
tb54      AGGTTATTAC TATTGGTGCT GAAAGGTTCC GTTGCCCGGA AGTTCTCTTC CAACCATCGA TGATCGGGAT
soy109    AGGTTATCAC AATTGGGGCA GAGAGATTCC GTTGCCCAGA AGTTCTTTTC CAGCCATCTA TGATTGGAAT
soy115    AGGTCATCAC CATTGGGTCT GAGCGTTTCA GATGTCCTGA AGTTCTCTTC CAGCCATCCA TGATCGGAAT
soy118    AGGTCATCAC CATTGGCGAT GAGCGTTTCA GACGTCCAGA GGTTCTGTTC CAACCATCCA TGATTGGAAT
soy119    AGGTCATCAC CATTGGTGCT GAGCGTTTCC GTTGTCCTGA AGTCCTCTTC CAGCCATCCA TGATTGGAAT
soy57     AGGTCATTAC AATAGGTTCT GAGAGATTCC GTTGCCCAGA AGTACTGTTC CAGCCTTCTC TGATTGGTAT
soy58     AAGTTATCAC CATAGGAGCT GAGAGATTCC GTTGCCCTGA AGTTCTTTTC CAGCCTTCTA TGATTGGAAT
soy69     AGGTTATCAC TATTGGCGCT GAGCGTTTCA GGTGTCCGGA GGTCTTGTAC CAACCATCCA TGGTAGGAAT
soy70     AGGTCATCAC CATTGGTGCT GAGCGTTTCA GATGTCCTGA AGTTCTCTTC CAACCATCCA TGATCGGAAT
soy86     AAGTTATCAC AATTGGGGCA GAGAGATTCC GTTGCCCAGA AGTTCTTTTC CAGCCATCTA TGATTGGAAT
SAc1      AGGTTATCAC TATTGGCGCT GAGCGTTTCG GATGTCCGGG GGTCTTGTAC CAACCATCCA TGGTAGGAAT
SAc3      AAGTTATCAC AATTGGGGCA GAGAGATTCC GTTGCCCAAA AATTCTTTTC CAACCATCTA TGATTGAAAT
mz56      AGGTTATCAC AATCGGCGCA GAAAGGTTTA GGTGCCCTGA GGTGCTATTC CAGCCATCTT TCATTGGCAT
mz63      AGGTTATCAC AATCGGCGCA GAAAGGTTTA GGTGCCCTGA GGTGCTATTC CAGCCATCTT TCATTGGCAT
mz81      AGGTGATCAC CATTGGGGCA GAGAGGTTCA GATGCCCTGA GGTCCTCTTC CAGCCTTCCT TCATTGGTAT
mz83      AGGTGATCAC CATTGGGGCG GAGAGGTTCC GATGCCCCGA GGTCCTCTTC CAGCCTTCCT TCATTGGTAT
mz87      AAGAATACAC CATCGGTGCA GAGAGATTCA GGTGCCCCGA GGTCCTCTTC CAGCCATCCA TGATTGGTAT
mz89      AGGTTATCAC AATCGGTGCA GAAAGGTTTA GGTGCCCTGA GGTTCTATTC CAGCCATCCT TCATTGGCAT
mz95      AGGTCATCAC GATCGGCGCG GAGAGGTTCA GATGCCCAGA GGTGCTTTTC CAGCCGTCGC TGATCGGGAT
MAc1      AGGTCATCAC CATTGGGTCA GAAAGGTTCA GGTGCCCCGA GGTGTTGTTC CAACCATCCC TTGTTGGCAT
```

Appendix: con't

							840
vlvx	GGAGGCGGTC	GGT---ATCC	ACGACACTAC	CTTTAACAGC	ATCATGAAGT	GCGATGTCGA	TATCCGCAAG
PsAc2	GGAAGCTGCT	GGA---ATTC	ATGAGACCAC	TTACAACTCT	ATAATGAAGT	GTGATGTTGA	TATCAGAAAG
PsAC1	GGAAGCTGCT	GGA---ATTC	ATGAGACCAC	TTACAACTCT	ATAATGAAGT	GTGATGTTGA	TATCAGAAAG
SoAc1	GGAAGCTGCT	GGC---ATTC	ACGAGACTAC	CTACAACTCC	ATCATGAAGT	GCGACGTGGA	TATTAGGAAG
AAc1	GGAGAATGCT	GGT---ATCC	ATGAAACCAC	CTATAACTCC	ATAATGAAGT	GTGATGTCGA	CATCAGGAAG
RAc1	GGAAGCTGCG	GGT---ATCC	ATGAGACTAC	ATACAACTCC	ATCATGAAGT	GCGACGTGGA	TATTAGGAAG
RAc3	GGAAGCTCCG	GGG---ATCC	ATGAGACCAC	CTACAACTCT	GTCATGAAGT	GTGAAGTGGA	TATCAGGAAG
RAc7	GGAAGCTCCT	GGC---ATCC	ATGAAGCCAC	ATACAACTCC	ATCATGAAGT	GCGATGTTGA	TATTAGAAAG
RAc2	GGAAGCTCCA	GGC---ATCC	ATGAGACGAC	ATACAACTCC	ATCATGAAGT	GCGATGTTGA	TATCAGAAAG
DcAc1	GGAGTCTGCT	GGA---ATCC	ATGAAACTAC	TTACAATTCC	ATTATGAAGT	GTGACGTTGA	TATCAGAAAG
pt46	GGAAGCTGCT	GGA---ATTC	ATGAGACGAC	GTATAATTCT	ATAATGAAGT	GTGATGTTGA	TATTAGAAAG
pt66	GGAAGCTGCT	GGT---ATTC	ATGAAACCAC	ATACAATTCC	ATCATGAAGT	GTGACGTTGA	TATCAGGAAG
pt79	GGAAGCTGCT	GGT---ATTC	ATGAAACCAC	ATACAATTCC	ATCATGAAGT	GTGACGTTGA	TATCAGGAAG
pt82	GGAAGCTGCT	GGA---ATTC	ATGAAACGAC	GTATAATTCA	ATCATGAAGT	GTGATGTGGA	TATTAGGAAG
PoAc101	GGAAGCTGCT	GGA---ATCC	ATGAAACAAC	TTACAATTCC	ATCATGAAGT	GCGATGTAGA	TATCAGAAAG
PoAc58	GGAAGCAGCT	GGA---ATCC	ACGAGACCAC	ATACAACTCG	ATCATGAAGT	GTGACGTTGA	TATCAGAAAG
PoAc71	GGAAGC?GCT	GGA---ATTC	ATGAGACGAC	ATATAATTCT	ATAATGAAGT	GTGATGTTGA	TATTAGAAAG
PoAc75	GGAAGCTGCT	GGT---ATCC	ATGAGACAAC	ATACAACTCT	ATAATGAAGT	GTGATGTTGA	TATCAGGAAG
PoAc97	GGAAGCTGCA	GGA---ATCC	ACGAGACTAC	ATACAACTCT	ATCATGAAAT	GTGATGTGGA	TATTAGAAAA
PoAc99	GGAAGCTGCA	GGT??AATCC	ATGAGACTAC	CTACAACTCC	TTTATGAAGT	GTGATGTTGA	TATCGGGAAG
pt42	GGAAGCTGCT	GGA---ATCC	ATGAAACAAC	CTACAACTCC	ATCATGAAGT	GTGACGTCGA	TATCAGGAAG
pt65	GGAAGCTGCT	GGA---ATCC	ATGAAACAAC	CTACAACTCC	ATCATGAAGT	GTGACGTCGA	TATCAGGAAG
tm32	GGAAGCTGCT	GGG---ATAC	ATGAAACAAC	GTACAACTCG	ATCATGAAAT	GTGACGTTGA	TATCAGGAAG
tm41	GGAAGCTGCA	GGA---ATCC	ACGAGACTAC	ATACAACTCT	ATCATGAAAT	GTGACGTGGA	TATTAGGAAA
tm52	GGAAGCTGCT	GGT---ATTC	ATGAAACCAC	GTACAATTCC	ATCATGAAGT	GTGACGTTGA	TATCAGGAAG
tm51	GGAAGCTGCA	GGT---ATCC	ACGAGACTAC	CTACAACTCT	ATTATGAAGT	GTGATGTTGA	TATCAGGAAG
tm105	GGAAGCAGTT	GGA---ATTC	ATGAAACAAC	ATATAGCTCA	ATCATGAAGT	GTGATGTGGA	TATTAGAAGA
tb103	GGAAGCTGCA	GGT---ATCC	ATGAGACTAC	CTACAACTCC	ATTATGAAGT	GTGACGTTGA	TATTAGGAAG
tb104	GGAAGCAGCA	GGA---ATTC	ATGAAACAAC	CTATAACTCA	ATCATGAAAA	GTGATGTGGA	CATTAGGAAG
tb66	GGAAGCTGCT	GGA---ATTC	ACGAAACAAC	TTACAATTCC	ATCATGAAGT	GTGATGTAGA	TATCAGAAAG
tb71	GGAAGCTGCA	GGA---ATCC	ACGAGACTAC	ATACAACTCT	ATCATGAAGT	GTGATGTGGA	TATTAGAAAG
tb93	GGAAGCTGCT	GGT---ATTC	ATGAAACCAC	ATACAATTCC	ATCATGAAGT	GTGACGTTGA	TATCAGAAAG
tob1	GGAAGCTGCA	GGC---ATTC	ACGAAACGAC	TTATAACTCG	ATCATGAAGT	GCGACGTGGA	TATCAGAAAG
tb53	GGAAGCAGTT	GGA---ATTC	ATGAAACAAC	CTACAACTCA	ATCATGAAGT	GTGATGTGGA	TATTAGGAAA
tb54	GGAAGCTGCA	GGT---ATCC	ATGAGACTAC	CTACAACTCC	ATCATGAAAC	GTGATGTGGA	TATTAGGAAG
soy109	GGAAGCTGCT	GGA---ATTC	ATGAGACCAC	CTACAATTCT	ATCATGAAGT	GTGATGTGGA	TATCAGAAGG
soy115	GGAAGCCGCT	GGC---ATTC	ATGAAACCAC	TTACAATTCA	ATCATGAAAT	GTGATGTCGA	TATTAGGAAG
soy118	GGAAGCTTCA	GGC---ATTC	ACGAGACAAC	ATACAACTCC	ATAATGAAGT	GCGATGTTGA	TATCAGGAAG
soy119	GGAAGCTGTT	GGC---ATTC	ATGAGACCAC	ATACAATTCA	ATTATGAAGT	GTGATGTTGA	TATCAGGAAG
soy57	GGAAGCTACT	GGA---ATTC	ATGAAACGAC	ATACAATTCC	ATCATGAAAT	GTGATGTTGA	TATTAGGAAG
soy58	GGAAGCTGCT	GGA---ATTC	ACGAGACCAC	CTACAACTCC	ATCATGAAGT	GTGATGTGGA	TATTAGGAAG
soy69	GGAAGCATCA	GGC---ATTC	ATGAAACGAC	ATACAACTCC	ATAATGAAAT	GTGATGTTGA	CATCAGGAAA
soy70	GGAAGCTTCT	GGC---ATTC	ATGAGACCAC	TTACAATTCA	ATCATGAAGT	GTGATGTCGA	CATTAGGAAG
soy86	GGAAGCTGCT	GGA---ATTC	ATGAGACCAC	CTACAATTCT	ATCATGAAGT	GTGATGTGGA	TATCAGAAAG
SAc1	GGAAGCATCA	GGC---ATTC	ATGAAACAAC	ATACAACTCC	ATAATGAAAT	GTGATGTTGA	CATCAGGAAA
SAc3	GGAAGCTGCT	GGA---ATTC	ATGAGACCAC	CTACAATTCT	ATCATGAAGT	GTGATGTGGA	TATCAGAAAG
mz56	GGAATCTCCT	GGC---ATCC	ATGAAGCCAC	GTACAACTCC	ATCATGAAGC	GTGACGTCGA	TATCAGAAAG
mz63	GGAATCTCCT	GGC---ATCC	ATGAAGCCAC	GTACAACTCC	ATCATGAAGT	GTGACGTCGA	TATCAGAAAG
mz81	GGAAGCTCCT	GGC---ATCC	ATGAGACCAC	CTACAACTCC	ATCATGAAGT	GCGATGTCGA	CATCAGGAAG
mz83	GGAAGCTCCT	GGC---ATCC	ATGAGACCAC	CTACAACTCC	ATCATGAAAT	GCGACGTGGA	TATCAGGAAG
mz87	GGAGGCTGCT	GGA---ATCC	ATGAGACAAC	ATACAATTCA	ATCATGAAGT	GTGATGTGGA	TATCAGGAAG
mz89	GGAATCTGCT	GGC---ATCC	ATGAAGCCAC	GTACAACTCC	ATCATGAAGT	GTGACGTCGA	TATCAGAAAG
mz95	GGAAGCTCCC	GGC---ATCC	ACGAGACCAC	CTACAACTCG	ATCATGAAAT	GTGACGTTGA	TATCAGGAAG
MAc1	GGAATCGCCT	AGC---GTCC	ATGAGGCCAC	GTACAACTCC	ATCATGAAGT	GTGATGTTGA	TATCAGGAAG

Appendix: con't

910

```
vlvx      GATCTGTACA ACAACATCGT GCTCTCTGGT GGAACGACCA TGTTCCCTGG CATCGCTGAT CGCATGAGCA
PsAc2     GATCTGTACG GAAACATTGT TCTTAGTGGT GGTACAACTA TGTTTCCCGG TATTGCTGAC CGTATGAGCA
PsAc1     GATCTGTACG GAAACATTGT TCTTAGTGGT GGTACAACTA TGTTTCCCGG TATTGCTGAC CGTATGAGCA
SoAc1     GATCTATATG GCAACATCGT CCTCTCTGGT GGTACCACTA TGTTCCCTGG GATTGCTGAC AGGATGAGGC
AAc1      GACTTGTACG GTAACATTGT GCTCAGTGGT GGAACCACAA TGTTCCCTGG AATCGCCGAC AGAATGAGCA
RAc1      GATCTATATG GCAACATCGT TCTCAGTGGT GGTACCACTA TGTTCCCTGG CATTGCTGAC AGGATGAGCA
RAc3      GATCTCTATG GTAACATTGT GCTTAGTGGA GGAACTACCA TGTTCCCTGG CATTGCTGAC AGGATGAGCA
RAc7      GATTTGTACG GTAATGTGGT CCTTAGTGGG GGATCTACAA TGTTCCCTGG TATTGGTGAT CGTATGAGCA
RAc2      GACCTGTATG GTAACATTGT TCTCAGTGGT GGATCCACCA TGTTCCCTGG CATCCGGAGA CCGATGAGCA
DcAc1     GATCTTTATG GTAATATTGT ACTTAGTGGG GGATCAACTA TGTTTCCTGG TATTGCTGAT CGTATGAGCA
pt46      GATTTGTATG GGAACATTGT GCTTAGCGGT GGGTCGACTA TGTTTCCTGG CATTGCTGAT AGAATGAGCA
pt66      GATCTATATG GAAACATTGT CCTCAGTGGT GGTACCACAA TGTTCCCTGG TATTGCTGAT AGGATGAGCA
pt79      GATCTGTATG GAAACATCGT CCTCAGTGGT GGTACCACAA TGTTCCCTGG TATTGCTGAT AGGATGAGCA
pt82      GATCTGTATG GAAACATTGT ACTTAGTGGT GGTACGACTA TGTTCAATGG AATTGCTGAT AGAATGAGCA
PoAc101   GATCTTTATG GTAACATTGT CCTCAGTGGT GGAACAACAA TGTTTCCTGG TATTGCTGAT CGTATGAGCA
PoAc58    GACCTCTACG GTAACATTGT TCTCAGTGGT GGTACTACCA TGTTCCCCGG TATTGCCGAT AGAATGAGCA
PoAc71    GATTTGTATG GGAACCATGT GCTTAGTGGT GGCTCGACTA TGTTTCCTGG CATTGCTGAT AGAATGAGCA
PoAc75    GATCTGTATG GGAACATTGT ACTTAGTGGT GGATCCACTA TGTTTCCTGG TATTGCTGAC AGAATGAGTA
PoAc97    GATCTTTATG GAAACATTGT GCTCAGTGGT GGTACTACCA TGTTCCCTGG TATTGCTGAT AGAATGAGCA
PoAc99    GACCTCTATG GAATACTTGT GCTCAGTGGT GGCTCAACCA TGTTCCCTGG TATTGTTGAT CGTATGAGCA
pt42      GATCTTTATG GAAACATTGT ACTTAGTGGC GGAACGACTA TGTTTCCTGG AATTGCTGAT AGAATGAGCA
pt65      GATCTTTATG GAAACATTGT ACTTAGTGGC GGAACCACTA TGTTTCCTGG AATTGCTGAT AGAATGAGCA
tm32      GATCTTTATG GTAACATTGT ACTTAGCGGG GGGACGACTA TGTTTCCAGG AATCGCTGAT AGAATGAGCA
tm41      GATCTTTATG GAAACATTGT GCTCAGTGGT GGTACTACCA TGTTCCCAGG TATTGCTGAT AGAATGAGCA
tm52      GATCTGTATG GAAACATCGT CCTCAGTGGT GGTACCACAA TGTTCCCTGG TATTGCTGAT AGGATGAGCA
tm51      GACCTCTACG GTAACATTGT GCTCAGTGGT GGTTCAACCA TGTTCCCTGG TATTGCTGAC CGTATGAGCA
tm105     GATTTATATG GAAACGTTGT GCTTAGTGGT GGCTCAACTA TGTTCCCTGG CATAACGGAT CGTATGAGCA
tb103     GACCTCTATG GCAACATTGT GCTCAGTGGT GGCTCAACCA TGTTCCCTGG TATTGCTGAT CGTATGAGCA
tb104     GATTTATACG GAAATATTGT ACTCAGTGGT GGCTCAACTA TGTTCCCTGG TATAGCAGAT CGAATGAGCA
tb66      GATCTTTATG GAAACATCGT CCTTAGTGGT GGAACAACAA TGTTTCCTGG CATTGCAGAT CGTATGAGCA
tb71      GACCTTTATG GAAACATTGT GCTCAGTGGT GGCTCTACCA TGTTCCCGGG TATCGCTGAT AGGATGAGCA
tb93      GATCTGTATG GAAACATTGT CCTCAGTGGT GGTACCACCA TGTTCCCGGG CATTGCTGAT AGGATGAGCA
tob1      GATCTGTATG GAAAGATTGT ACTCAGTGGC GGTTCGACTA TGTTTCCTGG AATTGCTGAT AGGATGAGTA
tb53      GATTTATATG GAAACGTTGT GCTTAGTGGT GGCTCAACTA TGTTTCCAGG GATAACAGAT CGTATGAGCA
tb54      GACCTCTATG GTAACATTGT TCTCAGTGGT GGCTCCACCA TGTTCCCTGG TATTGCTGAT CGGATGAGCA
soy109    GATCTATATG GCAACATTGT TCTTAGTGGT GGCTCAACTA TGTTTCCTGG TATTGCAGAC CGGATGAGCA
soy115    GACTTGTATG GAAACATCGT TCTCAGTGGT GGTTCTACCA TGTTCCCAGG TATTGCTGAC AGAATGAGCA
soy118    GACCTCTATG GTAACATTGT TCTTTCAGGA GGAACAACCA TGTTCCCTGG TATTGCTGAC AGAATGAGCA
soy119    GACTTGTATG GAAACATAGT CCTCAGTGGT GGTTCTACCA TGTTCCCTGG CACTGCTGAC AGAATGAGTA
soy57     GATCTGTATG GAAACATCGT GCTTAGTGGA GGGTCTACTA TGTTTCCGGG AATTGCTGAT CGCATGAGCA
soy58     GATCTGTATG GTAACATTGT TCTTAGTGGT GGCTCAACTA TGTTCCCTGG TATTGCTGAC CGTATGAGCA
soy69     GACCTGTACG GTAACATTGT CCTTTCAGGA GGTACAACCA TGTTCCCTGG CATCGCTGAT AGAATGAGCA
soy70     GACTTGTATG GAAACATTGT TCTCAGTGGT GGATCTACCA TGTTCCCAGG TATCGCTGAC AGAATGAGCA
soy86     GATCTATATG GCAACATTGT TCTTAGTGGT GGCTCAACTA TGTTTCCTGG TATTGCAGAC CGTATGAGCA
SAc1      GACCTGTACG GTAACATTGT CCTTTCAGGA GGTACAACCA TGTTTCCTAG CATCGCTGAT AGAATGAGCA
SAc3      GATCTATATG GCAACATTGT TCTTAGTGGT GGCTCAACTA TGTTTCTTGG TATTGCAGAC CGTATGAGCA
mz56      GACCTGTACG GTAATGTCGT CCTCAGTGGT GGCTCCACTA TGTTCCCTGG CATCGGTAAC CGTATGAGCA
mz63      GACCTGTACG GTAATGTCGT CCTCAGTGGT GGCTCCACTA TGTTCCCTGG CATCGGTGAC CGCATGAGCA
mz81      GACTTGTATG GTAACATTGT GCTCAGTGGT GGCACGACCA TGTTCCCTGG TATTGCGGAC CGTATGAGCA
mz83      GACTTGTATG GTAACATTGT TCTCAGTGGT GGCACAACCA TGTTCCCTGG CATCGCTGAC CGTATGAGCA
mz87      GACCTCTACG GCAACATTGT GCTCAATGGA GGTTCGACTA TGTTCCCCGG TATCGCTGAT CGTATGAGCA
mz89      GATCTGTACG GTAATGTTGT CCTCAGTGGT GGCTCCACTA TGTTCCCTGG TATTGGTGAC CGCATGAGCA
mz95      GACCTGTACG GGAACATCGT CCTCAGTGGA GGGACGACCA TGTTCCCTGG CATCGCTGAC AGGATGAGCA
MAc1      GATTTGTATG GTAACGTTGT CCTCAGTGGG GGTTTTACCA TGTTTCCTGG GATTGCCGAT CGTATGAGCA
```

Appendix: con't

```
                                                                                                                    980
vlvx       AGGAGATCAC GGCCCTTGCG CCGTCTGCCA TGAAGATCAA GGTCGTCGCC CCGCCGGAGC GCAAGTACTC
PsAc2      AGGAGATCAC TGCCCTGGCT CCTAGCAGCA TGAAGATTAA GGTTGTGGCT CCACCAGAGA GGAAGTACAG
PsAc1      AGGAGATCAC TGCCCTGGCT CCTAGCAGCA TGAAGATTAA GGTTGTGGCT CCACCAGAGA GGAGGTACAG
SoAc1      AAGGAAATCA CTGCCTTGCT CCTAGCAGCA TGAAGATCAA GGTGGTTGCT CCTCCAGAAA GGAAGTACAG
AAc1       AAGAGATCAC TGCTTTGGCT CCAAGCAGCA TGAAGATCAA AGTCGTTGCC CCTCCAGAGA GGAAATACTC
RAc1       AGGAAGATCA CTGCCTTGCT CCTAGCAGCA TGAAGATCAA GGTGGTCGCC CCTCCTGAAA GGAAGTACAG
RAc3       AGGAGGATTA CGCCCTTGCA CCAAGCAGCA TGAAGATCAA GGTCGTGGCA CCTCCTGAGA GAAAATACAG
RAc7       AGGAAATAAC TGCACTCTGC CCTGGCAGTA TGAAG----- -GTAGTTGCA CCGCCAGAGA GGAAATACAG
RAc2       AGGAAGATCA CCGCGCTGCG CCAAGCAGCA TGAAGATCAA GGTTGTCGCT CCGCCAGAGA GGAAGTACAG
DcAc1      AAGAAATCAC TGCTCTTGCC CCCAGTAGCA TGAAGATCAA GGTGGTGGCC CCTCCAGAAA GAAAGTACAG
pt46       AGGAGATTCA AGCCTTAGCA CCAAGTAGCA TGAAGATCAA AGTGGTTGCT CCTCCTGAAA GGAAGTATAG
pt66       AGGAAATTAC TGCATTAGCT CCTAGTAGCA TGAAGATCAA GGTTGTCGCC CCACCTGAGA GGAAGTACAG
pt79       AGGAAATTAC TGCATTAGCT CCTAGTAGCA TGAAGATCAA GGTTGTCGCC CCACCTGAGA GGAAGTACAG
pt82       AGGAAATTAC TGCATTGGCT CCAAGCAGCA TGAAAATTAA GGTTGTTGCT CCACCTGAAA GGAAGTATAG
PoAc101 AGGAAATTAC CGCTCTTGCT CCCAGCAGCA TGAAAATCAA GGTGGTGGCG CCTCCTGAAC GGAAGTACAG
PoAc58     AAGAACTTAC TGCATTGGCT CCAAGCAGCA TGAAGATTAA GGTTGTTGCA CCACCAGAAA GAAAATATAG
PoAc71     AGGAGATTCA AGCCTTAGCA CCAAGTAGCA TGAAGATCAA AGTGGTTGCC CCCCCTGAAA GGAAGTATAG
PoAc75     AGGAGATCAC CGCGTTAGCC CCTAGCAGCA TGAAGATTAA GGTTGTTGCA CCCCCTGAAA GGAAGTACAG
PoAc97     AAGAAATTAC TGCATTGGCT CCTAGCAGCA TGAAGATTAA GGTGGTCGCT CCACCAGAGA GGAAATACAG
PoAc99     AGGAAATCAC TGCCTTGGCT CCCAACAACA TGAAGATCAA GGTTGTTGCT CCACCGGAGG GAAAGTACAG
pt42       AAGAAATCAC AGCATTGGCT CCATGTAGCA TGAAAATTAA GGTCGTTGCA CCACCAGAAA GGAAGTATAG
pt65       AAGAAATCAC AGCATTGGCT CCAAGTAGCA TGAAAATTAA GGTCGTTGCA CCACCAGAAA GGAAGTATAG
tm32       AAGAAATCAC AGCATTGGCT CCAAGTAGCA TGAAAATTAA GGTGGTTGCA CCACCAGAGA GGAAGTATAG
tm41       AAGAAATTAC TGCATTGGCT CCTAGCAGCA TGAAGATTAA GGTGGTCGCT CCACCAGAGA GGAAATACAG
tm52       AGGAAATTAC TGCATTAGCT CCTAGTAGCA TGAAGATCAA GGTTGTCGCC CCACCTGAGA GGAAGTACAG
tm51       AGGAAATCAC CGCCTTGGCT CCCAGCAGCA TGAAGATTAA GGTTGTTGCT CCACCGGAGA GAAAGTACAG
tm105      AGGAATTCAC TGCTCTTGCT CCAAGCAGCA TGAAGATCAA AGTCGTTGCA CCACCTGAAA GCAAGTACAG
tb103      AGGAAATCAC CGCTTTGGCT CCCAGCAGCA TGAAGATTAA GGTTGTTGCT CCACCGGAGA GAAAGTACAG
tb104      AAGAAATTAC TGCTCTTGCT CCAAGCAGCA TGAAGATCAA GGTGGTTGCA CCACCCGAGA GGAAGTATAG
tb66       AGGAAATTAC TGCTCTTGCT CCCAGCAGCA TGAAGATCAA GGTGGTGGCG CCTCCTGAAC GCAAGTACAG
tb71       AAGAAATTAC TGCACTTGCT CCTAGCAGCA TGAAGATTAA GGTGGTCGCC CCACCAGAGA GGAAATACAG
tb93       AGGAAATTAC CGCATTAGCT CCAAGTAGTA TGAAGATCAA GGTTGTCGCC CCACCTGAGA GGAAGTACAG
tob1       AGGAAATTAC TGCCTTAGCT CCAAGCAGTA TGAAAATTAA GGTGGTAGCT CCACCTGAGA GGAAGTATAG
tb53       AGGAAATCAG TGCTCTTGCT CCAAACAGCA TGAAGATTAA GGTTGTTGCA CCACCTGAGA GGAAGTACAG
tb54      AGGAAATCAC TGCTTTAGCT CCAAGCAGCA TGAAGATTAA GGTTGTTGCG CCACCAGAGA GAAAGTACAG
soy109     AGGAGATCAC TGCTCTTGCT CCTAGCAGCA TGAAGATTAA GGTTGTAGCT CCACCAGAAA GAAAATATAG
soy115     AGGAAATCTC AGCTTCGGCC CCAAGCAGCA TGAAGATTAA GGTGGTTGCT CCACCTGAGA GAAAGTACAG
soy118     AGGAAATTTC GGCATTGGCA CCCAGCAGCA TGAAGATCAA GGTTGTAGCA CCACCTGAGA GAAAATACAG
soy119     AGGAAATCAC TGCTTTGGCC CCATGCAGCA TGAAGATCAG GGTGGTTGCA CGACCCGAGA GAAAGTACAG
soy57      AGGAAATTAG TGCTCTTGCC ACCAGTAGCA TGAAAATCAA GGTGGTTGCA CCTCCTGAGA GAAAGTACAG
soy58      AGGAGATCAC TGCACTTGCT CCTAGCAGCA TGAAGATTAA GGTTGTGGCT CCACCAGAGA GAAAGTATAG
soy69      AGGAAATCTC TGCTTTAGCC CCAAGCAGCA TGAAGATCAA GGTGGTAGCA CCTCCTGAGA GAAAGTATAG
soy70      AGGAAATCTC AGCTTTGG.C CCAAGCAGCA TGAAGATTAA GGTTGTTGCT CCACCTGAGA GAAAGTACAG
soy86      AGGAGATCAC TGCTCTTGC. CCTAGCAGCA TGAAGATTAA GGTTGTAGCT CCACCAGAAA GAAAATATAG
SAc1       AGGAAATTTC TGCTTTAGCC CCAAGTAGCA TGAAGATTAA GGTGGTAGCA CCTTCCGAGA GAAAGTTCGG
SAc3       AGGAGATCAC TGCTCTTGCT CCTAGCAGCA TGAAGATTAA GGTTGTAGCT CCACCAGAAA GAAAATATAG
mz56       AGGAGATCAC TGCCCTTGCG CCGAGCAGCA TGAAAATCAA GGTGGTGGCA CCGCCTGAGA GGAAATACAG
mz63       AGGAGATCGC TGCGCTTGCC CCCAGCAGCA TGAAGGTGAA GGTGGTTGCT CCGCCGGAGA GGAAGTACAG
mz81       AGGAGATCAC TGCCCTTGCG CCGAGCAGCA TGGAAATCAA GGTGGTGGCA CCGCCTGAGA GGAAATACAG
mz83       AGGAGATCAC CGCCCTTGCC CCGAGCAGCA TGAAAATCAA GGTGGTGGCT CCACCCGAGA GGAAATACAG
mz87       AGGAGATCAC TTCCCTTGCA CCAAGCAGCA TGAAGATCAA GGTCGTCGCA CCACCTGAGA GGAAGTACAG
mz89       AGGAGATTAC TGCACTTGCC CCCAGCAGCA TGAAGGTGAA GGTGGTTGCT CCGCCGGAGA GGAAGTACAG
mz95       AGGAGATTAC TGCCCTCGCG CCGAGCAGCA TGAAGATCAA GGTCGTGGCG CCTCCTGAGA GAAAATACAG
MAc1       AGGAGATCAC GTCCCTAGTT CCTAGCAGCA TGAAGGTTAA AGTAGTTGCG CCACCTAGAA GGAAGTACAG
```

Appendix: con't

```
                                                                                                  1032
vlvx      TGTG---TGG ATTGGTGGTT CAATCCTGGC GTCCTTGTCG ACCTTCCAGC AA
PsAc2     TGTC---TGG ATTGGAGGAT CAATCCTTGC ATCTCTCAGC ACCTTCCAGC AG
PsAc1     TGTC---TGG ATTGGAGGAT CAATCCTTGC ATCTCTCAGC ACCTTCCAGC AG
SoAc1     TGTC---TGG ATTGGAGGAT CCATCTTGGC ATCTCTCAGC ACATTCCAGC AG
AAc1      TGTC---TGG ATTGGAGGGT CCATCTTGGC CTCCCTCAGT ACCTTACAGC AG
RAc1      TGTC---TGG ATTGGAGGAT CCATCTTGGC ATCTCTCAGC ACATTCCAGC AG
RAc3      TGTC---TGG ATTGGAGGTT CTATCCTTGC CTCCCTCAGC ACATTCCAAC AG
RAc7      TGTC---TGG ATTGGTGGGT TCATCCTGGC TTCTCTCAGT ACCTTTCAGC AG
RAc2      TGTC---TGG ATTGGAGGAT CCATCTTGGC ATCTCTGAGC ACCTTCCAGC AG
DcAc1     CGATCTCTGG ATTGGAGGAT CGATCCTGGC TTCTCTCAGT ACCTTCCAGC AG
pt46      TGTC---TGG ATTGGAGGTT CAATTTTGGC ATCCCTCAGT ACTTTCCAAC AG
pt66      TGTC---TGG ATCGGAGGAT CTATCTTGGC CTCCCTCAGC ACATTCCAGC AG
pt79      TGTC---TGG ATCGGAGGAT CTATATTGGC CTCCCTCAGC ACATTCCAGC AG
pt82      TGTT---TGG ATTGGAGGCT CTATTTTAGC TTCTCTAAGT ACCTTCCAAC AG
PoAc101   TGTC---TGG ATTGGAGGGT CTATCTTAGC TTCTCTCAGC ACCTTCCAAC AG
PoAc58    TGTC---TGG ATTGGAGGTT CTATTTTGGC TTCCCTCAGC ACCTTCCAGC AG
PoAc71    TGTT---TGG ATTGGAGGTT CAATTTTGGC ATCCCTCAGT ACCTTCCAAC AG
PoAc75    TGTG---TGG ATTGGAGGTT CCATCTTGGC ATCCCTCAGT ACCTTCCAGC AG
PoAc97    TGTC---TGG ATTGGAGGCT CTATCTTGGC TTCCCTCAGC ACCTTCCAGC AG
PoAc99    TGTC---TTG ATTGGAGGAT CCATCCTAGC GTCAGTTAGC ACCTTCCAGC AG
pt42      TGTT---TGG ATTGGAGGTT CAATCTTGGC TTCTTTAAGT ACCTTCCAAC AG
pt65      TGTT---TGG ATTGGAGGTT CAATCTTGGC TTCTTTAAGT ACCTTCCAAC TG
tm32      TGTT---TGG ATTGGAGGTT CAATTTTGGC TTCTTTAAGT ACCTTCCAAC AG
tm41      TGTC---TGG ATTGGAGGCT CTATCTTGGC TTCCCTCAGC ACCTTCCAGC AG
tm52      TGTC---TGG ATCGGAGGAT CTATTTTGGC CTCTCTCAGC ACATTCCAGC AG
tm51      TGTC---TGG ATTGGAGGAT CCATCCTTGC ATCACTTAGC ACCTTCCAGC AG
tm105     CGTC---TGG ATTGGAGGAT CAATTCTAGC TTCCCTCAGT ACTTTCCAAC AA
tb103     TGTC---TGG ATCGGAGGAT CCATCCTCGC ATCCCTTAGC ACATTCCAGC AG
tb104     TGTC---TGG ATTGGAGGGT CAATTCTAGC TTCTCTCAGT ACTTTCCAAC AA
tb66      TGTC---TGG ATTGGAGGAT CTATCTTAGC TTCTCTCAGC ACCTTCCAAC AG
tb71      TGTC---TGG ATTGGAGGTT CTATCTTGGC TTCCCTCAGC ACCTTCCAGC AG
tb93      TGTT---TGG ATTGGAGGAT CTATTTTGGC CTCTCTCAGC ACATTCCAGC AG
tob1      TGTT---TGG ATTGGAGGTT CTATTTTGGC TTCTTTAAGT ACTTTCCAAC AG
tb53      CGTC---TGG ATTGGAGGAT CAATTCTAGC TTCCCTCAGT ACTTTCCAAC AA
tb54      TGTC---TGG ATTGGAGGAT CCATTCTTGC ATCCCTCAGC ACATTACAGC AG
soy109    TGTC---TGG ATTGGAGGAT CAATCCTTGC ATCCCTCAGC ACCTTCCAAC AG
soy115    TGTC---TGG ATTGGTGGCT CTATCTTGGC TTCTCTCAGC ACCTTCCAAC AG
soy118    TGTC---TGG ATCGGTGGTT CTATCTTGGC ATCTCTTAGC ACCTTCCAGC AG
soy119    TGTC---TGG ATTGGTGGCT CTATCTTGGC TTCTCTCAGC ACTTTCCAAC AG
soy57     TGTC---TGG ATTGGTGGTT CTATCTTGGC ATCACTTAGC ACCTTCCAAC AG
soy58     TGTC---TGG ACTGGAGGAT CAATCCTTGC GTCCCTCAGC ACCTTCCAGC AG
soy69     TGTC---TGG ATTGGAGGTT CTATCTTGGC ATCCCTGAGC ACCTTCCAAC AG
soy70     TGTC---TGG ATTGGTGGCT CTATCTTGGC TTCTCTCAGC ACTTTCCAAC AG
soy86     TGTC---TGG ATTGGAGGAT CAATCCTTGC ATCCCTCAGC ACCTTCCAGC AG
SAc1      TGTT---TGG ATTGGAGGAT CAATCCTTGC ATCTCTAAGC ACCTTCCAGC AG
SAc3      TGTC---TGG ATTGGAGGAT CAATCCTTGC ATCCCTCAGC ACCTTCCAGC AG.
mz56      TGTC---TGG ATAGGAGGAT CCATCCTTGC CTCCCTGAGC ACCTTCCAAC AG
mz63      CGTT---TGG ATTGGTGGCT CCATTCTGGC CTCTCTCAGT ACCTTCCAGC AG
mz81      TGTC---TGG ATAGGAGGAT CCATCCTTGC CTCCCTGAGC ACCTTCCAAC AG
mz83      TGTC---TGG ATAGGAGGGT CCATCCTTGC TTCCCTGAGC ACCTTCCAAC AG
mz87      TGTC---TGG ATAGGAGGGT CGATCCTAGC CTCGCTCAGC ACTTTCCAAC AG
mz89      TGTT---TGG ATTGGTGGTT CCATTCTGGC CTCCCTCAGT ACCTTTCAGC AG
mz95      TGTT---TGG ATTGGTGGCT CGATCCTTGC GTCGCTCAGT ACGTTCCAGC AG
MAc1      TGTC---TGG ATCGGTGGCT CTATTTTGGC TTCTCTCAGC ACTTTTCAGC AG
```

Appendix: con't