INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Université d'Ottawa • University of Ottawa

Université d'Ottawa • University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES
ET POSTDOCTORALES

FACULTY OF GRADUATE AND
POSTDOCTORAL STUDIES

JENSEN, Jennifer
AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M.Sc. (Biology)
GRADE - DEGREE

Department of Biology
FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

The Largest Subunit of RNA Polymerase II as a Molecular
Marker for Inferring Land Plant Phylogeny

Guy Drouin
DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

F. Chapleau	L. Gillespie
J. Vierula	

J.-M. De Koninck, Ph.D.
LE DOYEN DE LA FACULTÉ DES ÉTUDES
SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE
AND POSTDOCTORAL STUDIES

The largest subunit of RNA polymerase II as a molecular marker for inferring land plant phylogeny

JENNIFER JENSEN

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
in partial fulfillment of the requirements for the
M.Sc. degree in the
Ottawa-Carleton Institute of Biology

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

0-612-67824-5

Canada

Acknowledgements

Many thanks to all the people who have helped me fulfill this chapter in my life. In particular to Dr. Guy Drouin for providing the project idea and for his patience in explaining molecular evolution to me, to my committee members Dr. Lynn Gillespie and Dr. John Vierula for their time and advice, to my lab mates Jamie, Marianne, Mehrdad, Banoo, Claude and Hong Yan for their help, encouragement and especially for keeping science fun during the frustrating times, to Dr. Doug Johnson and Candace for their help and to Ed Bruggink at the Carleton University greenhouse for his generosity with plant tissues and for his identification of *Magnolia soulangiana*.

A very special thank you to Jamie and my family for believing in me and most importantly for their love and understanding. I am eternally grateful!

Table of Contents

Abstract

This is an initial study using the gene encoding the largest subunit of RNA polymerase II (RPB1) to elucidate evolutionary relationships among ten land plants. This is the first study to use a single nuclear protein-coding gene to examine seed plant evolution. Results show RPB1 to contain no base compositional bias and to evolve at a conservative rate that is similar in most species studied here. This gene also exists as a single copy in most species and contains enough phylogenetically informative sites to resolve all relationships among the seed plants in this study. Maximum parsimony, neighbor-joining and maximum likelihood analyses all generate identical tree topologies with similar support values at each node. The angiosperms are a monophyletic clade comprised of *Nymphaea* as the most basal angiosperm, followed by *Magnolia*, then *Arabidopsis* and a monophyletic monocot clade containing maize and *Oryza*. The gymnosperms also form a monophyletic clade with *Welwitschia* and pine grouped together and sister to a *Cycad* and *Zamia* clade. These findings concur with recent studies that refute the Anthophyte theory and place *Nymphaea* near the root of the angiosperm tree. The RPB1 sequence shows great promise to resolve the phylogenetic relationships among plants.

Introduction

The colonization of land by plants in the mid-Palaeozoic era over 400 million years ago (MYA) radically changed the earth's environment and set the stage for the evolution of life on land by other terrestrial organisms by reducing atmospheric CO_2 levels, which lowered the earth's temperature, and by increasing the levels of O_2 (Bateman *et al.*, 1998; Qiu & Palmer, 1999). The earth's constantly changing environment had considerable impacts on the physiological capabilities of land plants and was characterized by morphological and reproductive innovations that allowed early plants to inhabit a number of terrestrial locations (Kenrick & Crane, 1997; Renzaglia *et al.*, 2000). Such innovations included the two-phase life cycle between gametophyte and sporophyte, specialized sex organs, vascular and structural tissues, stomates for gas exchange, endomycorrhizal associations, and the evolution of the seed (Kenrick & Crane, 1997). This rapid radiation in plant diversity occurred over 30 MY and by the late Devonian period all extant major plant groups, except for the angiosperms, were recognizable (Thomas & Spicer, 1987).

The first plants:

The first plants, algae, inhabited marine environments and their plant body, known as a thallus, was mainly composed of photosynthesizing cells found singly or in intertwining filaments (Weier *et al.*, 1982). Algae are non-vascular plants, highly diverse in morphology and they are the primary producers in marine ecosystems (Weier *et al.*, 1982). Their typically small size and aquatic surroundings limited the need for vascular conducting tissues, roots, stems and leaves (Weier *et al.*, 1982).

It is firmly believed that the origin of land plants stemmed from a group of green algae known as the Charophytes (Mishler *et al.*, 1994). Land plants inherited elementary body plan characteristics from the charophycean algae that ultimately lead to their successful radiation, like the presence of cellulose in the cell wall, a multicellular body, a cytokinetic phragmoplast, plasmodesmata, a single-celled apical meristem with the ability to become branched, asymmetric cell division, reproductive and vegetative cell specialization, and finally, the development of a placenta (Graham *et al.*, 2000). In terms of plant systematics, there are inconsistencies regarding which order of the Charophytes is the sister group to land plants and whether the Charophytes are monophyletic or paraphyletic and these inconsistencies seem to be rooted in the highly diverse morphology of the algae (Mishler *et al.*, 1994; Bhattacharya & Medlin, 1998; Qiu & Palmer, 1999).

Land plant evolution:

Although there are unanswered questions concerning the dawning of land plant evolution, plant taxonomists are confident that land plants are monophyletic in origin and that bryophytes constitute the first land plants (Mishler *et al.*, 1994; Kenrick & Crane, 1997; Qiu & Palmer, 1999). Bryophytes inhabit moist or aquatic areas because, like algae, water is needed for fertilization (Weier *et al.*, 1982). Major differences between the green algae and bryophytes are that bryophytes are comprised of blocks or sheets of parenchyma tissue, a sterile jacket surrounds and protects the developing gametes, and bryophytes undergo an alternation of generations, where the sporophyte is dependent on the dominant gametophyte (Weier *et al.*, 1982). The direct exposure to the earth's atmosphere spurred the evolution of key characteristics in bryophytes such as cuticles, sporopollenin-walled spores and stomates, in order to deal with the new problem of desiccation (Bateman *et al.*, 1998). These adaptations to a terrestrial lifestyle allowed for a broader range of habitats to be occupied and laid the groundwork for a major diversification in land plant evolution (Bateman *et al.*, 1998). Bryophytes are divided into three distinct lineages, the liverworts, hornworts and mosses.

Molecular studies based on rDNA, the chloroplast *rbc*L gene, and three mitochondrial group II introns along with morphological and combined data sets all support the paraphyly of bryophytes (Mishler *et al.*, 1994; Lewis *et al.*, 1997; Qiu *et al.*, 1998). The analyses of the *rbc*L gene and the mitochondrial group II introns also reveal liverworts as the most basal lineage among the bryophytes and hornworts as the closest lineage to vascular plants (Lewis *et al.*, 1997; Qiu *et al.*, 1998). Conversely, a recent study by Renzaglia and associates (2000), based on a compiled data set of morphological, developmental and ultrastructural features, placed hornworts as the earliest land plants and a monophyletic moss and liverwort clade as the sister group to vascular plants. Consequently, it is quite evident that bryophytes and vascular plants are closely related, but the branching order among the bryophytes is still controversial and needs to be resolved for a better understanding of land plant evolution.

Living vascular plants can be divided into 10 divisions, where six of these divisions produce seeds and the remaining four divisions do not. The seed-producing vascular plants include four gymnosperm divisions (the Coniferophyta, the Cycadophyta, the Ginkgophyta and the Gnetophyta) and two angiosperm subdivisions (the Monocotyledonae and the Dicotyledonae) (Weier *et al.*, 1982). The four divisions that do not produce seeds, often called the lower vascular plants, include the Psilophyta, the Lycophyta (club mosses), the Sphenophyta (horsetails) and the Pterophyta (ferns) (Weier *et al.*, 1982).

Vascular plants refer to any plant with highly developed conducting tissues for the uptake and distribution of water, minerals, and nutrients. Vascular plants have a more erect body, in order to have a broader spore/seed dispersal range, and consequently, this lead to the evolution of conducting tissues (Thomas & Spicer, 1987). Other body plan characteristics include the development of a sporophytic apical meristem and the ability for the shoot apical meristem to proliferate, which permitted the production of organs and sporophyte branching, respectively (Graham *et al.*, 2000). Sporophyte branching granted increased body size, productivity, and reproductive potential along with the ability to continue growth, if any stem cells were destroyed or damaged (Graham *et al.*, 2000). This branching capacity also tipped the evolutionary scales toward a more dominant sporophyte body and a more reduced gametophyte body that is typical of higher plants (Graham *et al.*, 2000).

Lower vascular plants were very important in establishing the early terrestrial ecosystem as they evolved soon after the emergence of land plants and were much larger than bryophytes (Kenrick & Crane, 1997). These plants are found in moist, cool, shaded habitats and require free water for fertilization, just like the bryophytes (Weier *et al.*, 1982). Of the lower vascular plants, the pteridophytes represent the largest group with over 10 000 living species and they were among the dominant vegetation until the Mid-Cretaceous, when the flowering plants took over (Weier *et al.*, 1982; Thomas & Spicer, 1987). The evolution of the angiosperm herbs competed directly with the pteridophytes and this resulted in the natural adaptation of more competitive fern species during the Tertiary period (Thomas & Spicer, 1987). In addition, there are documented fern fossil deposits that cover the Lower/Middle Devonian boundary to the present (Rothwell, 1999).

The abundance of lower vascular plants in the fossil record had initially mislead scientists to believe that these plants comprised the first land plants, but now the abundance of fossil specimens are being used to elucidate the evolutionary relationships among this poorly resolved group (Qiu & Palmer, 1999). Psilophyta is comprised of two genera, *Psilotum* and *Tmesipteris*, but their placement varies from a basal position in lower vascular plants to a position nested within the ferns, due to two differing hypotheses about their simple sporophyte and gametophyte morphology (Doyle 1998; Rothwell, 1999; Kenrick, 2000; Soltis & Soltis, 2000). Firstly, the simplicity of these structures has been interpreted as plesiomorphic to the first vascular plants, thus placing them as the sister group to all other vascular plants, and secondly, these simple structures have been suggested to be a product of reduction, which places them higher in the tree as most closely related to the Filicales, one of the major fern groups (Doyle, 1998; Kenrick, 2000; Soltis & Soltis, 2000). This latter hypothesis is supported by four molecular data sets (Doyle, 1998). Past phylogenetic studies, based on both morphological and molecular data, have confirmed the monophyly of the Lycophyta and Psilophyta, but the monophyly of the Pteridophyta is doubtful (Bateman *et al.*, 1998; Qiu & Palmer, 1999; Rothwell, 2000; Kenrick, 2000). A study based on a 28 Kb inversion in the chloroplast genome that is lacking in lycopods and bryophytes, yet present in all other vascular plants, lends convincing evidence for a basal lineage of the lycopods among vascular plants; however, no other molecular study supports this hypothesis (Doyle, 1998; Qiu & Palmer, 1999). Presently, the resolution among the ferns and between the different groups of lower vascular plants is still quite poor and requires the addition of crucial information available in the fossil record for a clearer understanding of their evolutionary relationships.

The higher vascular plants, or seed plants, possess both vascular conducting tissues and seeds. Key innovations of this group include the seed, secondary growth and axillary branching (Doyle, 1998). The gymnosperm seeds lack a protective ovary wall and are said to be "naked", while the angiosperm seeds are borne within the enclosed ovary (Weier *et al.*, 1982). Secondary growth provided seed plants with an increased stability in their upright stems, it permitted the growth of some truly "arboreous forms", and it also granted the ability to adjust mechanical and conducting properties (Bateman *et al.*, 1998). On the one hand, it is believed that secondary growth arose before the evolution of the seed, due to a group of plants, known as the progymnospermophyta, that possess secondary growth, but reproduce by way of spores (Thomas & Spicer, 1987; Bateman *et al.*, 1998). On the other hand, there are a group of plants, known as the pteridospermophyta or seed ferns, that possess fern-like foliage, but reproduce by seeds suggesting that seeds evolved independently more than once (Thomas & Spicer, 1987; Bateman *et al.*, 1998). Both groups are intermediary between the lower vascular plants and the seed plants, but it is generally accepted that the progymnospermophyta are the ancestor of gymnosperms (Thomas & Spicer, 1987).

Gymnosperms dominated most of the earth's vegetation throughout the late Palaeozoic and Mesozoic, but have since undergone a decrease in number and distribution, due to the success of the angiosperms (Thomas & Spicer, 1987). As stated earlier, the extant gymnosperms are separated into four divisions the Coniferophyta, the Cycadophyta, the Ginkgophyta and the Gnetophyta. The conifers represent the most significant group of gymnosperms and they also have the oldest fossil record dating back to the Upper Carboniferous, approximately 310 MYA, with modern conifer families appearing sometime during the Triassic period (Thomas & Spicer, 1987; Miller, 1999). Cycads are a much smaller group with only nine genera and about 100 species with origins rooted somewhere in the Palaeozoic (Thomas & Spicer, 1987). The Ginkgophyta are represented by only one living species, *Ginkgo biloba*, and have a probable origin in the Upper Carboniferous (Thomas & Spicer, 1987; De Franceschi & Vozenin-Serra, 2000). The Gnetales refer to three highly specialized gymnosperms (*Ephedra, Gnetum* and *Welwitschia*) that have been the subject of one of the largest debates in seed plant phylogeny: Are the Gnetales more closely related to the angiosperms or the gymnosperms?

Morphological and molecular analyses suggested that the Gnetales, a monophyletic group with *Ephedra* as the most basal lineage and *Gnetum* and *Welwitschia* as a sister clade, were the closest living relatives of the angiosperms (Chase *et al.*, 1993; Doyle, 1996; Friedman, 1996; Price, 1996). This hypothesis is known as the Anthophyte Theory. However, two separate scientists have pointed out that previous molecular studies that supported the Anthophyte Theory were actually based on unrooted trees; therefore, the sister relationship between the Gnetales and the angiosperms could not have been inferred, due to a lack of outgroups that emerged prior to the evolution of seed plants (Frohlich, 1999; Samigullin *et al.*, 1999). Consequently, the Anthophyte Theory is solely supported by morphological data. An analysis by Chaw *et al.* (1997) using nuclear 18S rRNA sequences disagreed with the Anthophyte Theory and suggested that none of the extant gymnosperms were the ancestor of the angiosperms. Their study indicated that seed plants are a monophyletic clade and that the Gnetales are a sister group to the conifers. Recently, a number of research papers have been published that also disagree with the Anthophyte Theory (Hansen *et al.*, 1999; Samigullin *et al.*, 1999; Winter *et al.*, 1999; Bowe *et al.*, 2000; Chaw *et al.*, 2000; Donoghue & Doyle, 2000). Chaw (2000) proposed a new idea, the gnepines hypothesis, where the Gnetales are suggested to have arisen from within the conifers, based on the results that place the Gnetales as sister to either the Pinaceae or all conifers.

The monophyly of extant gymnosperms is readily accepted and Cycads followed by *Ginkgo* are the most basal lineages (Chaw *et al.*, 1997; Samigullin *et al.*, 1999; Bowe *et al.*, 2000; Chaw *et al.*, 2000). Since no living gymnosperm is closely related to the angiosperms, correct interpretations of morphological characters in the fossil record are necessary, if the correct relationships among seed plants are to be inferred. Whether or not all gymnosperms, extant and extinct, are a monophyletic group is still unresolved along with the monophyly of all seed plants (Chaw *et al.*, 1997). These issues need to be resolved in order to clarify the relationships amongst gymnosperms and also to determine if flowering plants are derived from an extinct gymnosperm lineage.

Angiosperms, or flowering plants, are the largest group with over 250 000 living species, which outnumbers all of the other land plant groups combined (Crane *et al.*, 1995). These plants are characterized by a number of features including enclosed ovules within carpels, the presence of vessels, reticulate leaf venation, a double integument, tectate pollen, the presence of a flower, double fertilization and lastly, phloem companion cells that are derived from sieve-tube mother cells (Thomas & Spicer, 1987). A major diversification of the flowering plants occurred in the mid-Cretaceous, between 130 and 90 MYA, and this marks the beginning of angiosperm domination, but many modern families did not evolve until the Tertiary period (Thomas & Spicer, 1987; Crane *et al.*, 1995). The success of the angiosperms is largely due to the many different types of floral forms that are specific to certain pollination vectors (Thomas & Spicer, 1987). Insect pollination played a significant role in the success of flowering plants because pollen could now be targeted for fertilization and outcrossing when plant individuals were geographically far apart (Thomas & Spicer, 1987). Insect pollination and scattered distribution by birds and animals increased the ability for angiosperms to survive and maintain genetic variability, particularly at low density (Thomas & Spicer, 1987). Flowering plants are still the most dominant vegetation today.

As previously mentioned, the angiosperms are separated into two subdivisions the Monocotyledonae and the Dicotyledonae. The monocots make up about 22% of the angiosperms, while the eudicots (non-magnoliid dicots) comprise about 75% and the basal magnoliid dicots make up the last 3% (Crane *et al.*, 1995). Monocots are nested within the basal dicots and appear to be monophyletic, but the dicots are clearly not monophyletic (Soltis & Soltis, 2000). The monocot-dicot divergence has been estimated, using molecular data, at approximately 200 MYA and based on phylogenetic dating, more than 50 but less than 100 angiosperm clades survived the mass extinction that marked the end of the Cretaceous period (Bremer, 2000).

The origin of the angiosperms, the other major debate in seed plant phylogeny, has eluded scientists for over a century and Charles Darwin called their sudden appearance an "abominable mystery". There are a number of hypotheses for the origin of extant angiosperms including the Magnoliales, Winteraceae, Calycanthales, Chloranthaceae, Ceratophyllaceae and Casuarinaceae hypotheses (Loconte, 1996). The first three all conform to the "ranalean theory" that states the order Magnoliales or the Winteraceae family as the most primitive angiosperms, while the fourth hypothesis suggests a shrub habit as primitive and the last two hypotheses suggest a paleoherb and a eudicot respectively as the most primitive angiosperms (Loconte, 1996). Interestingly, a Jurassic angiosperm fossil was found that showed similarities to the Magnoliales, however, this fossil was missing a perianth and this has been seen in some species of the Chloranthaceae family (Crepet, 1998). Consequently, this fossil, known as *Archaefructus*, has characters that support two competing theories for the origin of angiosperms and more importantly, it reveals a new subclass of angiosperms (Crepet, 1998).

It has been hypothesized that the first extant angiosperms were among the Magnoliales, Nymphaeales or the paleoherbs (herbaceous magnoliids + monocots) based on past morphological studies and molecular studies using 18S rDNA and *rbc*L sequences (Nickrent & Soltis, 1995; Soltis *et al.*, 1997; Doyle, 1998). Consequently, it was impossible to determine whether the first living angiosperms had a woody or herbaceous growth habit (Doyle, 1998). In 1999, breakthrough studies using duplicate phytochrome genes (Matthews & Donoghue, 1999) and multigene data sets from all three plant genomes (Parkinson *et al.*, 1999; Qiu *et al.*, 1999; Soltis *et al.*, 1999) surprisingly showed that a small, woody, shrub from New Caledonia, named *Amborella trichopoda* (A), was the root of extant angiosperms. Hence, a woody growth habit was more primitive and the herbaceous habit was derived from a woody ancestor. These recent studies also resolved the other basal lineages of angiosperms by identifying the second branch as the Nymphaeales (N) and the third branch as an Illiciaceae, Schisandraceae, Trimeniaceae and Austrobaileyaceae clade (ITA) (Matthews & Donoghue, 1999; Parkinson *et al.*, 1999; Qiu *et al.*, 1999; Soltis *et al.*, 1999(a); Kuzoff & Gasser, 2000). The three basal lineages are known as the "ANITA" clades. Barkman and colleagues

(2000) reanalyzed the data mentioned above and discovered that the most basal extant angiosperm lineage contained both *Amborella* and Nymphaeales. This finding changes how we look at the ancestor of angiosperms, but the inference of a woody nature remains unchanged (Barkman *et al.*, 2000). Similarly, Graham and Olmstead (2000) studied 17 chloroplast genes for inferring the relationships among the basal angiosperms. The ANITA clades were confirmed as the most basal lineages, but whether the root of extant angiosperms stemmed from an *Amborella* lineage or a Nymphaeales lineage was unresolved. Graham and Olmstead (2000) concluded that it was "premature to place confidence in the *Amborella* rooting of the angiosperms". Needless to say, the origin of angiosperms is still under debate.

The problems in plant phylogeny:

Although much has been learned about the relationships among plants over the past two centuries, it is apparent from this review that most plant relationships are controversial or unknown. Many of these problems can be explained by the following factors: firstly, the fossil record is incomplete and inadequate for most plants, secondly, there exists a large gap between the first land plants and their closest algal relatives, thirdly, many plant groups, including intermediate taxa, are extinct and can't be used in molecular studies, fourthly, the Siluro-Devonian radiation occurred so quickly that it is difficult to address evolutionary changes using molecular data, and lastly, the tremendous diversity between similar life cycle phases of many plant groups make it difficult to determine character homologies for morphclogical analyses (Bateman *et al.*, 1998;

Kenrick, 2000). Combined analyses of both molecular and morphological data, the addition of more taxa and characters, the use of genes from all three plant genomes and improvements to sequence analysis methods have all been suggested as ways to address the problems facing plant phylogeny (Graybeal 1998; Bremer *et al.*, 1999; Cummings *et al.*, 1999; Slowinski & Page, 1999; Barkman *et al.*, 2000).

The evolution of RNA polymerases:

RNA polymerases are responsible for the transcription of all cellular RNA. In prokaryotes there is one RNA polymerase, while in eukaryotes there are three distinct RNA polymerases. The eukaryotic RNA polymerases share the same enzymatic properties, yet they are distinctly specialized and regulated (Memet *et al.*, 1988). RNA polymerase I is responsible for the synthesis of all ribosomal RNA precursor molecules, RNA polymerase II transcribes all messenger RNA and some small nuclear RNA, while RNA polymerase III transcribes transfer RNA, 5S rRNA and all other RNA not synthesized by RNA polymerase II (Wolfe, 1993). The three eukaryotic RNA polymerases arose through duplication events of the ancestral prokaryotic RNA polymerase gene and diverged over time (Allison *et al.*, 1985). These duplication events occurred before the divergence of the major eukaryotic groups (Sidow & Thomas, 1994) and RNA polymerases II and III are more closely related to each other than either is related to RNA polymerase I (Memet *et al.*, 1988). RNA polymerase genes are ubiquitous, highly conserved, and they hold plenty of informative sites, which make them ideal candidates for marker molecules in phylogenetic studies (Puhler *et al.*, 1989; Klenk, *et al.*, 1994).

RNA polymerase genes were successfully used to elucidate early eukaryotic evolution (Zillig *et al.*, 1989; Puhler *et al.*, 1989; Iwabe *et al.*, 1991; Stiller *et al.*, 1998) along with relationships among plants (Denton *et al.*, 1998; Oxelman & Bremer, 2000), fungi (Liu *et al.*, 1999) and protists (Klenk *et al.*, 1995). In particular, the gene encoding the largest subunit of RNA polymerase II (RPB1) has provided useful insight into rhodophyte evolution (Stiller & Hall, 1998), protist evolution (Hirt *et al.*, 1999) and most recently, in the evolution of arthropods (Shultz & Regier, 2000).

The RPB1 gene is characterized by nine highly conserved domains, termed domains A through I, based on their similarity to bacterial and archaeal homologs (Allison *et al.*, 1985; Puhler *et al.*, 1989; Nawrath *et al.*, 1990). A schematic representation of the location and function of these domains is depicted in Figure 1. It seems that the largest subunit of RNA polymerase II plays a role in template binding because zinc finger motifs and potential helix motifs have been noted in the amino terminal portion of the gene (Palenik, 1992). Domain F may be involved in RNA chain translocation, since this domain contains a binding site for α-amanitin, which has been shown to inhibit transcription by blocking RNA chain elongation (Nawrath *et al.*, 1990; Palenik, 1992). Unique to RPB1 is the presence of domain I containing the XSPTSPS consensus sequence repeated a variable number of times (Nawrath *et al.*, 1990). These repeated units interact with transcription factors and serve as substrates for *cdc* 2 protein

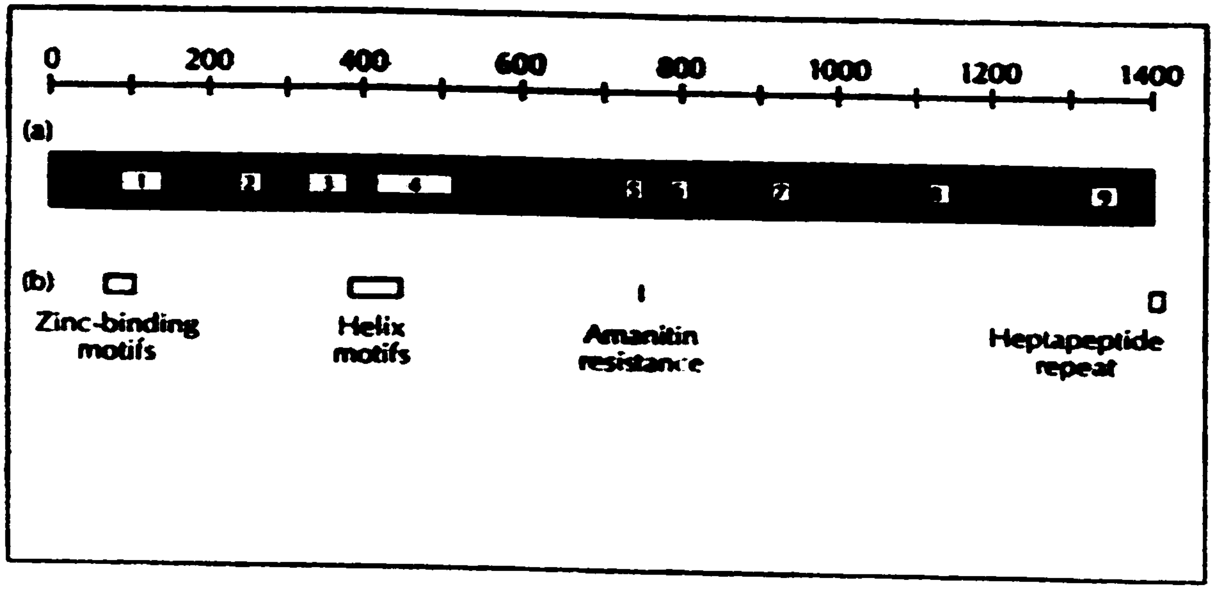


Figure 1 Schematic representation of the gene encoding the largest subunit of RNA polymerase II (RPB1). The nine conserved domains are numbered in white boxes (a) and possible functions are noted (b). These nine domains are most often termed domains A to I rather than domains 1 to 9 as shown in this figure. The scale corresponds to the amino acid sequence of the *Escherichia coli* β' subunit, which is homologous to the eukaryotic RPB1. Eukaryotic RPB1 has a similar structure, but it is usually larger. For example the *Arabidopsis thaliana* RPB1 sequence is 1742 amino acids long. Adapted from Palenik (1992) *Current Biology*.

kinases, which are involved in cell cycle control (Nawrath *et al.*, 1990). These repeats are also phosphorylated to varying degrees and their function is to ensure accurate transcription initiation of genes transcribed by RNA polymerase II (Palenik, 1992). The functions of the remaining domains are unknown. In addition, the genome sequencing projects for *Arabidopsis thaliana* and *Oryza sativa* have revealed the plant RPB1 gene to be approximately 5200 base pairs long. Due to the conserved nature of the RPB1 gene sequence, we were able to design primers specific to domains A and G. Consequently, we were able to amplify approximately 3.1 Kb of the RPB1 coding region and use this sequence information in phylogenetic studies of land plants.

This study illustrates the evolutionary relationships among ten land plants. Based on the past success of RPB1 in phylogenetic studies, we hypothesize that this gene will be a useful marker for elucidating the evolutionary relationships of land plants. In particular, we address the "gnepines" hypothesis proposed by Chaw (2000), which suggests that the Gnetales arose from within the conifers and we also address whether Nymphaeales or Magnoliales occupy the basal position in angiosperm phylogeny.

Materials and Methods

Plant Material:

The majority of plant tissues were collected from the University of Ottawa and the Carleton University greenhouses. Shehre-Banoo Malik, a former graduate student from this laboratory, collected the *Nymphaea* plant material from the Rideau River. Pine needles and *Ginkgo* leaves were collected during the spring and summer months from trees located on the University of Ottawa campus. The pine species was identified using a key developed by Dallimore & Jackson (1948). All tissue samples were washed and then stored until needed at -80°C. For a complete list of plant species see Table 1.

RNA/DNA Isolation:

All RNA and DNA extractions were from leaf or needle tissue. High quality genomic DNA for each plant species was isolated using the Qiagen Plant DNeasy Maxi Prep Kit as directed by the manufacturer. Each sample was eluted twice in 50 µl of sdH_20. A much higher yield of maize genomic DNA, with comparable levels of purity, was obtained using the Nucleon Phytopure Plant DNA Extraction kit (Amersham Pharmacia Biotech). With the exception of pine, all RNA was extracted using the Qiagen RNeasy Plant Mini Kit as per manufacturer's directions. The kit is supplied with two lysis buffers, Buffer RLT and Buffer RLC, which contain guanidinium isothiocyanate (GITC) and guanidine hydrochloride (GuHCL), respectively. Buffer RLT was chosen

Table 1. The A to G regions of RPB1 for the ten land plant species in this study.

Species	Common Names	Number of clones[a]	GenBank accession number[b]	Number of bases
Arabidopsis thaliana	Arabidopsis	N.A.	AL031986	3036
Oryza sativa	Rice	N.A.	AC084218	3030
Zea mays (Pioneer 3953)	Maize/Corn	4		3027
Nymphaea odorata	Water lily	3		3039
Magnolia soulangiana	Magnolia	5		3030
Zamia latifdia	Zamia	AD fragment 2 DG fragment 2		3039
Cycas revoluta	Cycad	5		3039
Pinus nigra	Pine	AD fragment 2 DG fragment 2		3039
Welwitschia mirabilis	Welwitschia	1		3039
Psilotum nudum	Whisk fern	4		3039

Note: The AD and DG fragments for pine and *Zamia* correspond to the overlapping PCR amplified fragments used in this study. All other RPB1 sequences were PCR amplified as one large fragment from domain A to domain G.

a, Multiple clones were obtained from more than one PCR sample.

b, We have not yet deposited our RPB1 sequences in GenBank.

N.A. = not applicable.

because of its greater cell disruption and denaturing properties (RNeasy manual, p. 45). The RNA was DNase treated according to the Qiagen RNase-Free DNase Set protocol. During the RNA isolation, the DNase treatment is applied directly to the spin column. The RNA samples were eluted twice in 50 μl of RNase free dH_20.

Pine RNA was successfully isolated following the Bahloul and Burkard (1993) protocol. All centrifugations were performed in a Sorvall Superspeed RC2-B centrifuge with an HB-4 rotor. DNase treatment was performed on the pine RNA samples using the Ambion DNA-free kit according to the directions of the manufacturer. The RNA integrity, purity and yield were all high when verified by gel electrophoresis and spectrophotometry.

cDNA synthesis:

DNase treated RNA samples were reverse transcribed into single-stranded cDNA by using Boehringer Mannheim's 1st Strand cDNA Synthesis Kit following the instructions of the manufacturer. The kit is supplied with Oligo-$p(dT)_{15}$ primer, which is used for priming the 3'-end of poly(A)-mRNA, and Random Primer $p(dN)_6$, which is used for priming nonspecific points along the RNA template. We were unable to produce a PCR product if the Oligo-$p(dT)_{15}$ primer was used, however, we were always able to produce the expected PCR fragment when we used the Random Primer $p(dN)_6$. All cDNA samples were then pooled together except for the negative control, which contained no AMV reverse transcriptase.

PCR amplification of RPB1:

An alignment based on *Arabidopsis thaliana*, *Glycine max*, and various animal and fungal sequences available on GenBank was used by Shehre-Banoo Malik and Dr. Guy Drouin to design degenerate primers to conserved amino acid motifs found within the nine domains of RPB1. The primers were designed to the ECPGHFG motif in domain A (5' GAI TGY CCI GGI CAY TTY GG 3' forward), to the MMVPKCIV motif in domain D (5' AC IAT RCA YTT IGG IAC CAT CAT 3' reverse), to the PYNADFDGDEMN motif also in domain D (5' CCI TAY AAY GCI GAY ITY GAY GGI GAY GAR ATG AA 3' forward), to the KAMVTAG motif in domain F (5' AAR GCI ATG GTI ACI GCI GG 3' forward), to the VGQQNVEG motif also in domain F (5' CC YTC IAC RTT YTG ICC IAC 3' reverse) and finally, to the GEPATQMT motif in domain G (5' GT CAT YTG IGT IGC IGG YTC TCC 3' reverse). Degenerate primers designed to domains A and G were used to amplify the RPB1 fragment of approximately 3.1 Kb in all plant species, except for pine and *Zamia*. Pine and *Zamia* were each amplified in two overlapping fragments using primers designed to domains A, D and G. The quantity of each primer used in the PCR experiments differed based on their level of degeneracy. For example, the primer designed to domain A contained three degenerate sites giving eight possible primers, while the primer designed to domain D contained only two degenerate sites giving four possible primers. Therefore, more of primer A was needed in the PCR experiments than primer D in order to account for its greater degeneracy.

RPB1 was amplified from the cDNA samples for most plant species under study by using Boehringer Mannheim's Expand High Fidelity PCR System according to the manufacturer's instructions. The Expand kit contains an enzyme mix of thermostable Taq DNA polymerase and Pwo DNA polymerase. According to the manufacturer, this allows for increased fidelity in DNA synthesis, due to the 3' - 5' proofreading activity of the Pwo DNA polymerase. The Expand kit is also designed for effectively amplifying longer PCR fragments. PCR experiments consisted of a Mg^{++} titration from 2.0 mM – 3.5 mM, a no template negative control, the cDNA negative control with template, a reaction involving the primers designed to domains A and D, and finally, a reaction involving the primers designed to domains F and G. These last two positive controls helped ensure full-length cDNA reactions. The concentration of the various reagents used was as indicated in the protocol included with the Expand High Fidelity PCR System kit. For every plant species, the PCR cycling conditions were as follows: 2 min. @ 94°C, (15 sec. @ 94°C, 30 sec. @ 45°C, 3 min. @ 68°C) cycled 10 times, (15 sec. @ 94°C, 30 sec. @ 50°C, 3 min. + Δ5 sec. @ 68°C) cycled 20 times, 7min.@ 72°C.

The *Zamia* AD fragment was PCR amplified using Taq polymerase (Amersham Pharmacia Biotech) under the following conditions: 2 min. @ 94°C, (40 sec. @ 94°C, 1 min. @ 55°C, 3 min. @ 74°C) cycled 35 times, 10 min. @ 74°C. The PCR experiment consisted of a no template negative control, the cDNA negative control with template, and four reactions composed of a (25mM) $MgCl_2$ titration between 2.0 mM and 3.5 mM, 8 µl (10 mM) dNTPs, 2 µl gelatin, 12 µl (5 µM) AD primer, 8 µl (5 µM) DA primer, 0.5

μl (5000 U/mL) Taq polymerase, 5 μl cDNA, 10 μl 10X buffer, and sdH_2O to a total volume of 100 μl.

Welwitschia DNA was amplified using the Accurase enzyme kit (Gordon Technologies), according to the manufacturer's specifications, and Hot Start PCR under the following cycling conditions: 2 min. @ 94°C, (20 sec. @ 94°C, 30 sec. @ 50°C, 2 min. @ 68°C) cycled 10 times, (20 sec. @ 94°C, 30 sec. @ 55°C, 2 min. + Δ20 sec. @ 68°C) cycled 20 times, 7 min. @ 68°C. The *Welwitschia* PCR also contained a 2.0 mM – 3.5 mM Mg^{++} titration and a no template negative control.

Mehrdad Hajibabaei, a graduate student in this laboratory, provided the *Zamia* and pine clones corresponding to the D to G regions.

Cloning and plasmid purification of the RPB1 PCR fragment:

Each PCR product corresponding to the desired fragment size was gel isolated and purified by using the UltraClean15 (MOBIO Laboratories) or the GEL ECLIPSE DNA Purification Kit (Gordon Technologies) according to the directions of the manufacturers. Each sample was eluted in 20 μl of sdH_2O. 3'A overhangs were then added to each of the gel purified fragments because the Expand High Fidelity PCR System generates a mixture of products containing 3'A overhangs and blunt ended fragments. This is a direct result of the 3' - 5' exonuclease activity of the Pwo DNA polymerase. Therefore, in a final volume of 20 μl, 2 μl of 10X PCR buffer (Amersham

Pharmacia Biotech) was added along with 0.5 µl (10 µmol) dATP and 0.5 µl (5000 U/mL) Taq polymerase (Amersham Pharmacia Biotech). The samples were then incubated at 72°C for 10 - 15 minutes and are now ready to be used in a cloning reaction.

Invitrogen's TOPO TA Cloning kit (vector pCR 2.1) was used as per the instructions of the manufacturer in order to obtain transformants containing the RPB1 insert. Prior to plating, the cell culture was spun for 1 minute at 13 000 rpm in a Biofuge pico microcentrifuge (Heraeus Instruments). 225 µl of the supernatant was removed and the pelleted cells were resuspended in the remaining 75 µl of cell culture media. This entire amount was then plated on Luria-Bertani (LB) agar plates containing ampicillin for antibiotic selection, along with IPTG and X-gal for colour selection of the transformants. Later versions of the TOPO TA Cloning kit (vector pCR 2.1) added a salt solution to the ligation reaction to increase efficiency and also no longer required the addition of IPTG to the LB agar plates.

The putative RPB1 clones were then PCR colony screened using the Accurase enzyme kit because of its superior performance over Taq polymerase with longer fragments. The protocol included with the Accurase enzyme kit was followed. However, the colony screen total reaction volume was reduced from 50 µl to a final reaction volume of 12.5 µl and we utilized the degenerate PCR amplification primers designed to the conserved motifs of RPB1 during the colony screen of the putative clones. The cycling conditions were as follows: 2 min. @ 94°C, (20 sec. @ 94°C, 30 sec. @ 50°C, 2 min. @ 68°C) cycled 10 times, (20 sec. @ 94°C, 30 sec. @ 53°C, 2 min. @ 68°C) cycled 20 times, 7 min. @ 68°C.

During the PCR colony screen set-up, the sterile toothpick used to pick the chosen colony from the LB plate was dipped in a PCR colony screen sample tube and then dipped in 500 µl of LB liquid media containing ampicillin. The liquid cultures were incubated at 37°C on a shaker set to 250 rpm for 3 – 4 hours. The liquid cultures corresponding to the positively screened RPB1 clones were then added to 5 mL of liquid LB containing ampicillin and grown overnight at 37°C, while shaking at approximately 250 rpm. The overnight cultures were used to retrieve purified plasmid preparations and to prepare plasmid stocks (1.3 mL of culture and 300 µl of 75% glycerol), which were subsequently stored at -80°C.

In order to obtain highly pure plasmid preparations for sequencing, the Qiagen QIAprep Spin Miniprep Kit was used according to the directions of the manufacturer. All samples were eluted twice in 50 µl of sdH_2O (i.e., the 50 µl of sdH_2O recovered from the first elution was loaded onto the column and eluted a second time). Restriction digests were performed in a total volume of 15 µl, so that we could establish which RPB1 putative clones were of the correct size to be sequenced. Depending on the plant species used, varying restriction enzymes were needed to drop the entire insert. For example, Eco RI will drop the *Welwitschia* RPB1 insert, but *Nymphaea* RPB1 has a number of Eco RI internal cut sites and had to be dropped using Hind III and Xho I restriction enzymes.

Sequencing of the putative RPB1 clones:

Sequencing was performed as directed by the Perkin-Elmer Big Dye Terminator Sequencing kit instructions and the excess dye terminators were removed according to the CENTRI-SEP protocol (Princeton Separations). The University of Ottawa Biotechnology Institute carried out further processing of the reactions. As of March 2001, we processed the sequencing reactions ourselves on an ABI 310 Genetic Analyzer by means of capillary electrophoresis.

Internal primers needed to be designed in order to obtain the full-length sequence of both DNA strands. The internal primers were designed from consensus sequence alignments for each plant in the study. The majority of primers were degenerate and due to the varying melting temperatures (Tm) of the internal sequencing primers, the annealing temperature of the sequencing reactions ranged from 43°C to 50°C (Appendix 1).

Southern blot hybridizations:

Southern blot analyses were performed to determine the RPB1 gene copy number in the plant species studied here. 10 µg aliquots of plant genomic DNA were 40X over-digested by three different AT-rich restriction enzymes, Eco RI, Xba I and Bgl II, respectively, for 4 hours at 37°C. The 400 µl digests were phenol/chloroform extracted and precipitated with ethanol. The digests were then washed with 70% ethanol, resuspended in 50 μl of TE (pH 8) and heated to 65°C for 7 minutes and quenched on ice for 5 minutes. The genomic DNA digests were run on a 0.7% SeaKem GTG agarose gel in 1X TAE and electrophoresed overnight at 34V. The agarose gel was depurinated in a 0.5M HCL solution for 10 minutes, denatured twice using a solution of 1.5M NaCl, 0.5M NaOH for 30 minutes and neutralized twice using a solution of 1.5M NaCl, 1M Tris (pH 8) for 45 minutes. Hybond-N nylon membranes (Amersham Pharmacia Biotech) were used in the overnight nucleic acid blotting transfers. The following morning, the nylon membranes were rinsed in 2X SSC, allowed to dry at room temperature and were UV cross-linked in a CL-1000 Ultraviolet cross linker. The nylon membranes were then stored at -20°C in a 50 mL falcon tube until needed.

The DNA probes we used were an amplified PCR fragment from the plant mini preps corresponding to domains F and G. These PCR fragments were purified using columns from the QIAGEN QIAquick PCR Purification Kit as directed by the manufacturer. A serial dilution of the DNA probes varying in DNA concentration from 1 ng to 1 *f*g in 6X SSC buffer were prepared and denatured at 95°C for 10 minutes and quenched on ice for 5 minutes. This serial dilution was spotted on Hybond-N nylon membrane and it acted as a control dot blot for each plant probe. The dot blots were then denatured in a solution of 1.5M NaCl, 0.5M NaOH for 5 minutes, neutralized in a solution of 1.5M NaCl, 1M Tris (pH 8) for 2 minutes, rinsed in a solution of 2X SSC for 5 minutes and allowed to dry at room temperature for 2 minutes. The control dot blots were UV cross-linked in a CL-1000 Ultraviolet cross linker. The control dot blots were stored in a 50 mL falcon tube at -20°C until needed.

The hybridization buffer contained 5X SSC, 5% Denhardt's, 0.5% SDS, 100 μg/mL of sheared calf thymus DNA, and sdH_2O. The Southern blots and control dot blots were prehybridized in 20 mL of hybridization buffer at 65°C for a minimum of 1 hour. The DNA probes were then labeled according to the manufacturer's instructions for the T7 Quick Prime Kit (Amersham Pharmacia Biotech). Probe Quant G-50 Micro Columns (Amersham Pharmacia Biotech) were utilized to purify the ^{32}P α dCTP labeled DNA probes. The probes were then added to the hybridization tubes containing 20 mL of fresh buffer and hybridized overnight at 65°C.

Moderate stringency washes were performed on the Southern hybridization blots. For *Welwitschia, Psilotum, Cycad* and maize the wash conditions were as follows: 2X SSC, 0.1% SDS twice at room temperature, 2X SSC, 0.1% SDS three times at 65°C, and 1X SSC, 0.1% SDS once at 65°C with gentle shaking for 10 minutes at each wash step. This was repeated for *Magnolia* and *Nymphaea* along with an additional wash of 1X SSC, 0.1% SDS at 65°C with gentle shaking for 10 minutes. The *Zamia* blot was washed in 2X SSC, 0.1% SDS twice at room temperature and once at 65°C with gentle shaking for 10 minutes at each wash step. The Southern blots were then exposed to BioMax MS Kodak film enveloped in a Kodak BioMax TranScreen-HE intensifying screen at -80°C for 5 – 10 days. The film was then placed in developer for 3 minutes, rinsed in H_2O, placed in fixative for 3 minutes, rinsed in H_20 and left to dry.

Sequence analysis:

All sequences were verified using the National Center for Biotechnology Information (NCBI) Blast search web page located at www.ncbi.nlm.nih.gov/BLAST/. The translating blast search (blastx) option was chosen as it converts a nucleotide sequence query into the six protein reading frames and then compares these translated sequences against the protein databases found at NCBI. The default settings for a non-redundant database and a standard genetic code were used during the query.

The plant RPB1 sequences were assembled into contigs and consensus sequences using the Sequencher (version 4.0.5) software from Gene Codes Corporation. These consensus sequences were then manually edited using the Genetic Data Environment (GDE) version 2.2.

Sequences were aligned using Clustal W version 1.8 (Thompson *et al.*, 1994). Default settings were applied except for the output format option, which was changed from the CLUSTAL format to the PHYLIP format. The sequences corresponding to the primers used in the PCR amplification were deleted from the beginning and the end of each plant RPB1 sequence prior to the alignment. The DNA alignment was then compared to the amino acid alignment and minor adjustments were made in GDE, so that the nucleotide alignment matched the protein alignment. The nucleotide alignment was then used to calculate the base composition, dinucleotide composition and codon usage for each of the ten taxa by the CODONS version 1.4 program (Lloyd & Sharp, 1992). The codon usage table and G + C content option were chosen from the available codon usage choices. The Li93 program (Li, 1993) was used to determine the number of synonymous (Ks) and nonsynonymous (Ka) substitutions per site between the 10 plant sequences. This analysis was performed using a weighted path, since RPB1 is a functional gene, and we excluded gaps. To establish if the rate of nonsynonymous nucleotide substitutions per site evolved at significantly different rates between the lineages, the RRTree program version 1.1.7 (Robinson *et al.*, 1998) was implemented. *Psilotum* was the chosen outgroup and the amino acid alignment was used as the input file. The Bonferroni correction (Holm 1979) was applied to deal with statistical artifacts created by multiple comparisons.

When reconstructing the phylogenetic trees, the PHYLIP package version 3.5c (Felsenstein, 1993), PUZZLE version 4.0.2 (Strimmer & von Haeseler, 1996), and TreeView version 1.2 (Page, 1996) were used. All analyses were based on the amino acid alignment and rooted with *Psilotum*. The default settings of PROTPARS were employed to search for the best protein parsimony tree, except that the sequences were jumbled once and a different seed number was utilized in each of the ten trials. The total number of parsimony informative sites was determined in MEGA version 1.02 (Kumar *et al.*, 1993). The protein distance matrix was generated in PROTDIST with its default settings and the Dayhoff PAM matrix model. The best neighbor-joining tree was found by randomizing the input order of the sequences in ten independent runs of NEIGHBOR. Bootstrap analyses were performed by generating 100 replicates in SEQBOOT and the

CONSENSE tree program was used to obtain the majority-rule consensus parsimony and distance trees. Maximum likelihood analyses of the amino acid sequences were implemented by the PUZZLE program using exact quartet likelihood, exact parameter estimates, the JTT model of substitution and a model of rate heterogeneity composed of one invariable and eight gamma rates estimated from the data set. Rooting of the trees with *Psilotum* and minor editing was completed in TreeView.

To test alternative tree topologies, the Kishino-Hasegawa test was implemented by the PUZZLE program using user defined trees, exact parameter estimates based on the 1st input tree (best tree), the JTT model of substitution and a model of rate heterogeneity composed of one invariable and eight gamma rates estimated from the data set. Tested hypotheses included the ranalean theory (i.e., *Magnolia* at the base of the angiosperms), the Gnetales as sister to the Cycads, the Gnetales placed within the Cycads, the Gnetales as sister to all gymnosperms and finally, the Anthophyte Theory (i.e., the Gnetales as the closest extant relatives to the angiosperms).

Results

RNA Isolation from gymnosperm tissues:

Several protocols were tried to isolate RNA from pine, *Araucaria* and *Ginkgo*, but the majority gave pitiful RNA integrity and/or very low yields and purity. The following is a list of the ineffectual methods and which plants were attempted with each one: TRIzol Reagent supplied by GIBCO BRL (*Araucaria*); Qiagen RNeasy Plant Kit including suggested amendments (pine, *Araucaria*, *Ginkgo*); Straight A's mRNA Isolation System supplied by Novagen (pine); Hot Borate Method (Hamby *et al.*, 1988) (pine); low pH method (Bahloul & Burkard, 1993) (*Araucaria, Ginkgo*); Doug Johnson's GITC method based on a modified Chomcyzynski & Sacchi protocol (1987) (*Araucaria, Ginkgo*); no chaotropic salt method (Bugos *et al.*, 1995) including suggested amendments (pine, *Ginkgo*). Consequently, an effective RNA isolation protocol for *Araucaria* and *Ginkgo* was never found. Another attempt at the Bahloul & Burkard (1993) method should be made because the Oakridge centrifuge tubes we used may have been contaminated with RNases; in addition, this is the method that was successfully used in the isolation of Pine RNA.

Sequence alignments:

The amino acid (Fig. 2) and nucleotide (Fig. 3) sequence alignments of the RPB1 gene began in domain A and terminated in domain G. The alignments for the ten plant species in this study consisted of 1014 amino acids and 3042 base pairs. The alignments

```
                            Domain A
ZAMIA           HLELAKPMYH IGFMKTVLSI LRSVCFDCSR ILADE-EDHR FKQALKIKNP KHRLRKVLDC
CYCAD           ....V..... ...L...... ......N... .......... .......... ..........
PINE            ..X....... ...L...... ..C...N... ......D... ......V... ....K.....
WELWITSCHIA     ........F. .......... ..C...N... .......... .......... ....E.I...
PSILOTUM        ........F. ..VL....T. ..C..YN... .......... ....R..R.. .....RM...
NYMPHAEA        ........F. ........A. ..C...N..H .......... .......R.. .A..K.....
MAGNOLIA        ........F. ...L...... M.C...N..K ..V......K ...SQ..R.. .........A
ORYZA           ........F. ...I...... M.C...N..K ......D.IK .......R.. .NK.KRIY.A
MAIZE           ........F. ...I...... M.C...N..K ......DETK .......R.. .N..KRIY.A
ARABIDOPSIS     Y......... V......... M.C...N..K .....VCRSL .R..M..... .N..K.I..A
```

```
ZAMIA           CKNKVKCEGG DEIEDDQAQD GEETAKKKRH GGCGAQQPKI SIDGMKIIAE YKATRKKADE
CYCAD           ....T..... ...D...... ...MT..... .......... .......... ..........
PINE            ....T..... .V.DE..G.. .D.EV..... .......... T......... ...P......
WELWITSCHIA     ....T..... .V.....G.. VD.EV...P. ....S..... T......... ...P......
PSILOTUM        ..T.S..... .D.DE..GH. N..AE...K. .......... T......... F..SK..NE.
NYMPHAEA        ..S.Q..... .D.DE..G.. L.DAP...KR .....M.... T.E....... ..VP...S..
MAGNOLIA        ....T..D.. .D..-VRD.. L..PV..R.- .........L TVE...M... .......N.D
ORYZA           ...RKI.A.. .NLD-V.E.Q .TDDPV...- ........N. TV....MV.. ...PK..N.D
MAIZE           ..S.KV.A.. .DLD-V.E.. -TDEPI...- ........N. TV....MV.. F..PK..T.D
ARABIDOPSIS     ....T..D.. .D.D.V.SHS TD.PV..S.- .........L T.E...M... ..IQ...N..
```

```
                                                                  Domain B
ZAMIA           QEQLMPEPVE RKQQLSAEKV LDILKRISDE DCQLLGLNPK YARPDWMILQ VLPIPSPPVR
CYCAD           .......... .......... .......... ...M...... .......... .....P....
PINE            ........A. ........R. .NV...T... E..M...... .......... ...V.P....
WELWITSCHIA     .......... ........R. .N........ E..M..F... .......... .....P....
PSILOTUM        .D.Y...... .......... .S...Q.... ..VT...... .......... .....P.S..
NYMPHAEA        ...ML..... ........R. .NV....... ..L....... .......... .....P....
MAGNOLIA        ....-..... ........R. .SV....... ..L....... .......... .....P....
ORYZA           ....-....D ...I....R. .NV..H.... ..L....... F......... .....P....
MAIZE           .D..-..... ...I....R. .NV....... ..L....... .......... ...V.P....
ARABIDOPSIS     PD..-...A. ...T.G.DR. .SV......A ......F... F........E .....P....
```

```
                            Domain B
ZAMIA           PSVMMNTTAR SEDDLTHQLA MIIRHNENLR RQERNGAPAH IITEFAQLLQ FHIATYFDND
CYCAD           .....D..S. .......... .......... .......... .......... ..........
PINE            .....D..S. .......... .........K K..K...... .......... .........E
WELWITSCHIA     ......S.T. .......... .........K K..K...... .V........ ..V......E
PSILOTUM        .....D.S.. .......... ......N..K ...Q...T.. ..N....... ..........
NYMPHAEA        .....D.SS. .......... ........X. .......... ..S....... .........E
MAGNOLIA        .....D.SS. .......... .......... .......... ..S....... .........E
ORYZA           .....D.SS. .......... .......... .......... .......... .........E
MAIZE           .....D.SS. .......... .......... .......... .......... ..........
ARABIDOPSIS     .....DA.S. .......... .........K ...K...... ..S..T.... .........E
```

```
                                                   Domain C
ZAMIA           LPGQPRATQR SGRPIKSICN RLKAKEGRIR GNLMGKRVDF SARTVITPDP NINIDQLGVP
CYCAD           .......... .......... .......... .......... .......... .....E....
PINE            .......... .........S .......... .......... .......... ..........
WELWITSCHIA     .......... .........S .......... .......... .......... .....E....
PSILOTUM        .......... .........Q .......... .......... .......... .....E....
NYMPHAEA        .......... .........S .......... .......... .......... T....E....
MAGNOLIA        .......... .........S .......... .......... .......... T....E....
ORYZA           .......... .........S .......... .......... .......... .....E....
MAIZE           .......... .........S .......... .......... .......... .....E....
ARABIDOPSIS     .........K .........S .......... .......... .......... T....E....
```

```
                   Domain C
ZAMIA         WSIALNLTYP ETVTPYNIER LKELVDNGPH PPPGKTGARY IIRDDGQRLD LRYLKKSSDH
CYCAD         .......... .......... .......... .......... .......... ..........
PINE          .......... .......M.. .....EY... ........K. ...E...... ..........
WELWITSCHIA   .......... .I........ .....EY... ........K. .......... .....R....
PSILOTUM      .......... .......... .....EY... .......... ...E...... ..F......R
NYMPHAEA      ......M... .......... .....E.... ........K. ...E....V. ....R.....
MAGNOLIA      .......... .......... .....EY... ........K. ...E...... ..........
ORYZA         .......... .......... .....EY... ........K. ...E...... ...V.....Q
MAIZE         .......... .......... .....EY... .X......K. ...E...... ...V.....Q
ARABIDOPSIS   .......... .......... ......Y... ........K. .......... .........Q

                                                      Domain D
ZAMIA         HLELGYKVER HLNDGDFVLF NRQPSLHKMS IMGHRIKIMP YSTFRLNLSV TSPYNADFDG
CYCAD         .......... ..M....... .......... ......R... .......... ..........
PINE          .......... .......... .......... .......... .......... ..........
WELWITSCHIA   .......... ......L... .......... .......... .......... ..........
PSILOTUM      .......... .......... .......... ....K.R... F......... ..........
NYMPHAEA      .......... ..I....... .......... .......... .......... ..........
MAGNOLIA      .......... .......... .......... .......... .......... ..........
ORYZA         .......... .......... .......... .......... .......... ..........
MAIZE         .......... .......... .......... .......... .......... ..........
ARABIDOPSIS   .......... ..Q....... .......... ......R... .......... ..........

                 Domain D
ZAMIA         DEMNMHVPQS FETRAEVLEL MMVPKCIVSP QSNRPVMGIV QDTLLGCRKI TKRDTFIEKD
CYCAD         .......... .......... .......... .......... .......... ..........
PINE          .......... .......... .......... .......... .......... ..........
WELWITSCHIA   .......... .......M.. .......... .......... .........V ..G...T...
PSILOTUM      .......... .......... .......... ......I... .........V ..........
NYMPHAEA      .........C .......... .......... .......... .......... ..........
MAGNOLIA      .......... .......... .......... .......... .......... ..........
ORYZA         .......... .......... .......... .......... .......... .....L....
MAIZE         .......... .......... .......... .......... .......... .....L....
ARABIDOPSIS   .......... .......... .......... .A........ .......... ..........

                                            Domain E
ZAMIA         VFMNILMWWE DFDGKIPXPT ILKPRPLWTG KQVFNLIIPR QINLIRYSAW HXESETGFIT
CYCAD         .......... N......S.. .......... .......... .......... .S........
PINE          .......... .......A.. .Y....I... .........K ......T... .S...S.V..
WELWITSCHIA   .......... E.E....N.A ......I... .........K .......... .N...K..P.
PSILOTUM      .......... .......A.. ....K..... ..L.S....K ......F... .LD..V.E..
NYMPHAEA      .......... .......A.A .M........ .........K .......... .S........
MAGNOLIA      .......... .......A.. .......... .........K ......T... .S.A......
ORYZA         .......... .....V.A.A ......I... .........K ......F.G. .S.A..R...
MAIZE         .........Q .......A.. ......I... .........K ......F... .S.E.K....
ARABIDOPSIS   ....T..... .....V.A.A .......... .........K ....L..... .ADT......

ZAMIA         PGDTVVRIEK GEVLSGTLCK KTLGTSSGSL IHVIWEEVGP DAARKFLGHT QWLVNYWLLQ
CYCAD         ....C..... .......... .......... .......... .......... ..........
PINE          ....L..... ...I...... .......... .......... .......... ..........
WELWITSCHIA   .......... .......... ....S..... .......... .......... ..........
PSILOTUM      .......... ..LI...... ......G... .......... .......... ..........
NYMPHAEA      ....C....R ..L....... ..M....... .......... .......... ..........
MAGNOLIA      ....Q....R ..L.A..... ......T... .......... .......... ..........
ORYZA         ....M..... ..L....... ......T... .......... .......... ..........
MAIZE         ....M..... ..L....... .S...G.... .......... .......... ..........
ARABIDOPSIS   ....Q....R ..L.A..... ......N... V......... .......... ..........
```

```
ZAMIA          QGFSIGIGDT IADAATMETI NETISKAKNE VNHLIQLAHQ KALEAXPGRT MMESFENRVN
CYCAD          .......... .......... ........A. ..Q....... .....E.... ..........
PINE           .......... ........V. .......... .KM..KA.QE .Q...E.... ..........
WELWITSCHIA    .......... ........K. ....AN..IK .QQ..EK.QE .K...E.... .......Q..
PSILOTUM       H......... ........K. ..S.A..... .KE..KM.QE .Q...E.... T.........
NYMPHAEA       NA........ ....S...K. .......... .KE..KA.QE .Q...E.... ..........
MAGNOLIA       N......... ....S...K. .......... .KE..KA.QE .Q...E.... ..........
ORYZA          N......... ........N. .........D .KK..KQFRD NQ...EA... T.........
MAIZE          N......... ....S..... .D.......A .KE..KK..E .Q...E.... ..........
ARABIDOPSIS    N..T...... ...SS...K. .....N..TA .KD..RQFQG .E.DPE.... .RDT......

                                                               Domain F
ZAMIA          QVLNKARDDA GSSAQKSLSE SNNLKAMVTA GSKGSFINIS QMTACVGQQN VEGKRIPFGL
CYCAD          .......... .......... .......... .......... .......... .........F
PINE           .......... .......... .......... .......... .......... .......Y.F
WELWITSCHIA    .......... .N........ .......... .......... .......... .......Y.F
PSILOTUM       ....R..... .....R.... ...V...... .......... ..I....... .........F
NYMPHAEA       .......... .....R.... .......... .......... .......... .........F
MAGNOLIA       .......... .......... .......... .......... .......... .......Y.F
ORYZA          E.......V. ....E..... .......A.. ....T..... .......... .........F
MAIZE          .......... .....N.... .......... .......... .......... .........F
ARABIDOPSIS    .......... ........A. T......... .......... .......... .........F

                                          Domain F
ZAMIA          FGRSLPHFTR DDYGPESRGF VENSYLRGLT PQEFFFHAMG GREGLIDTAV KTSETGYIQR
CYCAD          .......... .......... .......... .......... .......... ..........
PINE           D..T.....K ..K....... .......... .......... .......... ..........
WELWITSCHIA    N..T.....K ..K....... .......... .......... .......... ..........
PSILOTUM       VD.T.....K .......... .......... .......... .......... ..........
NYMPHAEA       ID.T.....K .......... .......... .......... .......... ..........
MAGNOLIA       ID.T.....K .......... .......... .......... .......... ..........
ORYZA          TN.T.....K N......... .......... .......... .......... ..........
MAIZE          ID.T.....K .......... .......... .......... .......... ..........
ARABIDOPSIS    D..T.....K .......... .......... .......... .......... ..........

                 Domain F
ZAMIA          RLVKAMEDIM VKYDGTVRNS LGDVIQFLYG EDGMDAVWIE SQKLDSLKMK RPEFENVYRY
CYCAD          .......... .......... .......... .......... .......... .S........
PINE           .......... .......... .......... .......... .......... KK......K.
WELWITSCHIA    .......... .......... ....T..... .......... .......... KR..DD..K.
PSILOTUM       ........V. .......... .....E.... .......... ......M... TK..DRM.MF
NYMPHAEA       .......... .......... .......... .......... .......... KS..DK.FKF
MAGNOLIA       .......... .......... .......... .......... .......... KR..D..F..
ORYZA          .......... .......... .......... ......I... .......... KA..D..F..
MAIZE          .......... .......... .......... .......... .......... K...D..F..
ARABIDOPSIS    .......... .......... .......... .......... .......... KS..DRTFK.

ZAMIA          EIDNENWNPS YMLPEHVEDL KTIREFQNVF DAEVQKLEAD RRQLGTEIAP SGDNSWPMPV
CYCAD          .......... .......... .......... ..Q....... .......... ..........
PINE           .L.Q...... .......... .......... ........S. .....M.... .......L..
WELWITSCHIA    .L.Q.....G ......I... .........C .....H..N. .......... .......L..
PSILOTUM       ...S.....D ...A...... ...Q..R... .......... ....C..... .......L..
NYMPHAEA       .F.D...D.. .LQ.DF.... ...K.L.D.. H.......T. ....A....T T.........
MAGNOLIA       .F..A..... .......... ......R... .......... .Y.......T T.........
ORYZA          .L.D...K.N .LSTQ.A... ...S.IR... E......... .F.......T T...T.....
MAIZE          .L.D...R.N .......D.. ......R... E......... .Y...S..TT T.........
ARABIDOPSIS    ...D.....T .LSD..L... .G...LRD.. ...YS...T. .F.......T N..ST..L..
```

```
ZAMIA        NLKRLIWNAQ KTFKIDLRKP SDMHPMEIVE AVDKLQERLK VVPGDDLMSI EAQKNATLFF
CYCAD        .......... .......... .......... .......... .......... ..........
PINE         ........S. ......A.R. .......... .I.R...... .......... ..........
WELWITSCHIA  .......... .L....P..A ........ID .I........ ....N.PL.. ..........
PSILOTUM     .......... .I..V.F... .......... .......... ....E.EI.K ..........
NYMPHAEA     .......... ....V...R. .......... .......... ....E.A..M ..........
MAGNOLIA     .......... ....V...R. .......... .......... ......FL.M ..........
ORYZA        .......... ........R. .........D .I........ ......DI.. ..........
MAIZE        .......... ......F.R. .......... .I........ ......A... ..........
ARABIDOPSIS  .I..H..... .........I .....V...D .........L ......AL.V ..........

                                                   Domain G
ZAMIA        NILLRSTFAS KRVLKEYRLT REAFEWVIGE IESRFLQSLV APGEMIGCVA AQSI
CYCAD        .......... .......... .......... .......... .......... ....
PINE         .......... .......H.. .......V.. .......... .......... ....
WELWITSCHIA  .......... .......H.. K...D..... ..T....... S......... ....
PSILOTUM     .......L.. .......... K......... ..T.....I. .......... ....
NYMPHAEA     .......... ...S...... ....D..... .......... .A........ ....
MAGNOLIA     .......... .......... .......... .......... .......... ....
ORYZA        .......... .......... K......... .......... .......... ....
MAIZE        .......... .......... K......... .......... .......... ....
ARABIDOPSIS  .......L.. ....E..K.S .......... .......... .......... ....
```

Figure 2 The RPB1 amino acid sequence alignment of the ten plant species used in this study. All amino acids identical to the *Zamia* sequence are represented by a dot and all dashes correspond to gaps in the alignment. The conserved A to G domains are labeled and shaded. The domains were identified according to the study of Nawrath *et al.* (1990).

```
ZAMIA          CACCTGGAAT TGGCAAAACC AATGTATCAT ATCGGCTTTA TGAAGACAGT TCTGTCGATT
CYCAD          .....T.... ...T...... C.....C... .........C .......C.. ..........
PINE           .......WG. .A........ T......... .........T .......C.. ......A...
WELWITSCHIA    ..T..T.... ....G..... G....T.... ..A..G.... ....A..T.. ...A..T...
PSILOTUM       ...T.....C .T..C..G.. G....T...C ..T..TG..C .......T.. ...CA.A...
NYMPHAEA       ........GC .C..C..G.. G....TC... .....G..C. .......G.. G..TG.T..C
MAGNOLIA       .....T..GC .C..C..G.. C....TC... .....G..CT ....A..T.. G..C..T..C
ORYZA          .....T..GC .T..T..G.. .....TC..C ........C. .C.....T.. A..C..C..A
MAIZE          .....T..GC .T..C..G.. G....TC..C .....A..C. .T.....T.. G..C..C..C
ARABIDOPSIS    T.T..T..GC .C..T..G.. .......... G....T.... .......... GT.AAGT..C

ZAMIA          TTGCGCTCAG TTTGTTTCGA CTGCTCGCGA ATCCTTGCCG ATGAG---GA GGATCATCGT
CYCAD          ........T. .G..C..TA. ......C..T .......... .C........ ..........
PINE           C.T..G.GC. ........A. T........C ..T..C..A. .......... T..C......
WELWITSCHIA    C....A.GT. ........A. .........T ..TT.G..T. .C........ ..........
PSILOTUM       C.A..G.GT. ....C.A.A. T..T.....C ..TT.G..T. .......... A........C
NYMPHAEA       C......GC. .C..C..TA. T.......AC ..T..C.... .......... ...C.....G
MAGNOLIA       A......GC. .C..C..TA. T.....TAA. ..A....TT. .......... A..C..CAAA
ORYZA          A......GC. .C..C...A. T.....CAAG ........A. .......... T...ATCAAG
MAIZE          A....T.GT. ....C..TA. ......CAAG .....C..G. .......... T..AACCAAA
ARABIDOPSIS    A..A.A.GT. .C......A. T.....CAAG ..TT.A..T. .....GTATG TAGGAGCTTA

ZAMIA          TTCAAACAAG CATTGAAAAT CAAGAACCCA AAGCACAGGC TTAGAAAGGT TTTGGATTGT
CYCAD          .....G.... .......... .........G .......... .......... ..........
PINE           .......... ....A...G. ......T..C ..A......T .A.A...... ..........
WELWITSCHIA    .......... .C..A..... A..A..T..G ..A..T..AT .GGAG...A. .C.T.....C
PSILOTUM       .....G.... ..AGA..G.. .CGC.....C ..A..TC..T .A....G.A. G..A......
NYMPHAEA       .....G.... .C........ ..GA..T... ..AGCTC.A. .G.AG..A.. .........C
MAGNOLIA       .....G...T .TCA...... T.GA..T... ..A..TC.A. .C..G..... .......GCA
ORYZA          .....G..G. .TC.T..... ..G...T... ..AA.T.AA. .A.A..GAA. A.AT...GC.
MAIZE          ..T..G..G. .T..A..... .CG...T... ...A.T..AT .G.AG.G.A. A.AT...GC.
ARABIDOPSIS    ..TCGT..G. .TA....G.. ......T..T ...A.T.... ...AG...A. .C.....GCC

ZAMIA          TGTAAAAACA AAGTAAAATG TGAGGGTGGA GATGAGATTG AGGATGATCA AGCTCAAGAT
CYCAD          .......... ..AC...... .......... .......... .T........ ..........
PINE           .....G..T. ..AC...... .......... ....T..... .T..G..... ..GA......
WELWITSCHIA    ........T. ..ACT..... .......... ....TA..A. .......... ..G...G..C
PSILOTUM       .....G.CT. .GTC...G.. C.....A..C .....C.... .T..A..... ..G...C...
NYMPHAEA       .......G.. .GCA...... ...A.....T .....C..C. .T..G..... ..GA..G...
MAGNOLIA       .....G..T. ..AC...... ...T.....T .....T.... ..---.TC.G ..A.......
ORYZA          ..C..G.... G.AAG.TC.. ..CA..C..T ..CA.TC... .T---.T... G.AG...C.A
MAIZE          ..C..G.G.. ..AA.GTC.. ..CA.....C .....CC... .T---.T... G.AA..G...
ARABIDOPSIS    ..C....... .GACC..... ...T.....T .....C.... .C....TC.. .AGC..CAGC

ZAMIA          GGGGAAGAGA CAGCAAAGAA GAAGCGTCAT GGAGGTTGCG GAGCTCAGCA GCCAAAAATT
CYCAD          .......... TGA....... .......... ........T. ....A..A.. A.........
PINE           ..C..T..AG AG.TT..... .......... ........T. .T....A... A..C..G..A
WELWITSCHIA    .TT..T...G AG.T...... .....C.... .....C..T. .TT.A..... A..C.....C
PSILOTUM       AAT..G...G ...A...A.. ...AAAA..C ........T. .T..A..A.. ...T.....A
NYMPHAEA       CTA.....CG .TC....... ...AAAA.G. ..G..C..T. .T...AT... A........A
MAGNOLIA       TTA.....AC .G.T...... .CGC..A--- ..G.....T. .T....A... ......GC.C
ORYZA          ..TACT..TG ATC..GT... ...AA.G--- ..T..C..T. .T........ ...G..T..C
MAIZE          ---ACT..TG AGC.T.TT.. ...AA.A--- ..T..C..T. .T........ ......T...
ARABIDOPSIS    AC...T..AC ...T...A.. ..GC..A--- ..T..A..T. .T..A..A.. A......C.G
```

```
ZAMIA        TCCATTGATG GCATGAAAAT AATTGCCGAG TATAAAGCTA CACGCAAGAA AGCTGATGAA
CYCAD        .......... .......... ......G... .......... .......... ...C.....G
PINE         A......... .T........ ......T..A .........C .......... ...A.....G
WELWITSCHIA  A....C.... .G........ T.....A... .........C ....G..A.. G..A..C..G
PSILOTUM     A.......C. .......G.. ......A... .T...G..C. GCAAA..... .AA...A...
NYMPHAEA     A.A.....A. .......G.. ......A..A .....G.T.C .CA.G..... .T........
MAGNOLIA     A.TG....A. .T........ G.....A..A ..C..G.... ..A.G..A.. GAA......T
ORYZA        A.TG....C. .T.....G.. GG....A... ........AC ..AAG..... .AAC.....C
MAIZE        A.AG.C.... .......G.. GG....T... .T.......C ..AAG..... .A.......T
ARABIDOPSIS  A.T.....G. .T.....G.. G.....A..A ..C...AT.C A.A.G..... .AA......G

ZAMIA        CAAGAACAGT TAATGCCAGA GCCTGTTGAG AGGAAACAAC AGCTTAGTGC AGAAAAGGTC
CYCAD        ........AC .......G.. ......G... ..A....... .A........ G.........
PINE         .........C .......... .....CA..A .......... .....TCA.. ...GCG...T
WELWITSCHIA  .........C .G........ ......G... ........G. .A..CTC... .....G...G
PSILOTUM     ..G..T.... AT.....T.. A.....C... ..A.....G. .A........ ...G.....T
NYMPHAEA     ..G..G..AA .GC....C.. A..A.....A ..A....... ....ATC... T....G...T
MAGNOLIA     .........C .T---..T.. A.....G..A ..A....... .....TC... .....G...T
ORYZA        .....G..AC .C---..T.. A.....G..C C....G...A TC..CTC... T..G.G....
MAIZE        ..G..T..AC .C---..T.. ...A..G..A C.C..G...A TC..CTCA.. T..G.G....
ARABIDOPSIS  .C...T...C .T---..C.. .....CA..A ........GA CA...G.... T..T.G...T

ZAMIA        TTGGATATTT TGAAGCGAAT AAGTGACGAA GATTGTCAAT TGTTGGGCTT GAACCCCAAA
CYCAD        .......... ....A..... ...C..T... .........A .......... ..........
PINE         C.AA..G... .......... ..C...T... ..G..C...A .......T.. ..........
WELWITSCHIA  C..A...... .......... T.....T... ..A......A ..C.T..G.. C.....A..G
PSILOTUM     C.AAGC..CC ....A.A... T.....T... ......GT.A CCC.T..... ...T..T..G
NYMPHAEA     C..A..G.CC .......... T..C..T..G .....CTT.C .TC....G.. ......G..G
MAGNOLIA     C.AAG.G..C .C.....G.. ......T..G ..C..C.T.C .......T.. ...T..T..G
ORYZA        C.TA..G..C .T..A.AT.. .........G ..C....TTC .A.....TC. ..........
MAIZE        C.TA..G..C .T.....T.. ......T..G ..C....TTC ........C. ...T......
ARABIDOPSIS  ...AG.G... ....AA.G.. T.......CG .........C .CC.A..T.. C.....T..G

ZAMIA        TACGCTCGCC CAGATTGGAT GATCTTACAG GTCCTTCCAA TTCCTTCACC TCCAGTTAGG
CYCAD        ..T.....T. .......... ...T..G..A .......... ....CC.T.. ..........
PINE         ..T.....A. .......... ...T..G..A ..T..C..TG .....C.T.. ...T..C..A
WELWITSCHIA  ..T..C..A. .......... ......G..A ..A.....T. .....C.... ...T.....A
PSILOTUM     ..T..A..A. .T..C..... ...AC....A ..TT.G..C. ....AC.T.. .T....G...
NYMPHAEA     ..T.....T. .T........ ...T..G... ........C. ....GC.T.. ...T..G...
MAGNOLIA     ..T..C..T. .G........ ...T.....A .....C..G. ....CC.T.. ...T..G..A
ORYZA        .TT..C..T. .T........ ...T..G..A ........G. .....C.... C..T..G..A
MAIZE        ..T..C..T. .T........ ...AC.T..A ..T.....TG .....C.... C..T.....A
ARABIDOPSIS  .TT.....T. .T..C..... ...TC.TG.A ........T. .....C.... C..T..C..A

ZAMIA        CCTTCAGTCA TGATGAACAC TACTGCCAGG AGTGAGGATG ATCTTACTCA TCAGCTAGCT
CYCAD        .....T.... .....G.... ....T.A... .......... ....C..... ..........
PINE         .....T.... .....G.... ...AT.A... .......... ....C..A.. ......T..G
WELWITSCHIA  ..C..T.... .......T.G ....A.A... .......... .......A.. ......T..A
PSILOTUM     ..C..T..T. .....G.... AT....AC.A .....A.... .C..C..... ...A..G..A
NYMPHAEA     ..G..T..G. .....G.T.. .T..T.AC.C .......... ..T.A..... ...A..G..C
MAGNOLIA     .....T..G. .....G.T.. AT..T.T... .......... .CT.G..... ....T.G..A
ORYZA        ..A..T..A. .....G.T.. .T..T....A .......... ..T.G..... ...AT....A
MAIZE        ..A..T.... .....G.... CT..T....A .......... ....G..... ....T.....
ARABIDOPSIS  ..A..T..A. .....G..G. C...T..... .......... .CT.G..C.. ..........
```

```
ZAMIA         ATGATAATTA GGCATAATGA AAATCTGAGG AGGCAGGAGA GAAATGGAGC TCCGGCTCAC
CYCAD         .....C.... ....C..C.. .......... .......... .......... ...A......
PINE          .....T..C. .A........ ........A. .A......A. AG.....G.. ...A.....T
WELWITSCHIA   .......... .A..C..C.. ....T...AA .A...A.... AG........ ...T......
PSILOTUM      .....T..CC .T..C...A. T..CT...AA C.C..A...C AG........ ...AA.A...
NYMPHAEA      .....C.... .A..C..... G.....Y... ........A. .......... ...T..A..T
MAGNOLIA      .....T.... .A..C..C.. G........A .......... .......... ...T..C...
ORYZA         .........C .A..C..... G..CT..... .....A.... .......... C..A......
MAIZE         ........AC .A........ G...T..... .....A.... .......T.. ...A......
ARABIDOPSIS   .....T...C .A..C..... ...CT...AA ........A. A......... G..A.....T

ZAMIA         ATCATCACTG AATTTGCTCA ATTGCTGCAG TTCCACATAG CTACCTATTT TGATAATGAC
CYCAD         ..T....... .......C.. .......... ........T. .......... ......C..T
PINE          ..T..T.... .......... ....T....A ........T. .......... ...C.....G
WELWITSCHIA   ..TG.T..G. .G.....A.. G..A.....A ..T..TG.T. .A........ ...C.....G
PSILOTUM      ..T..T.A.. .......C.. GC.AT..... ........T. .C........ ..........
NYMPHAEA      ..A..TT.A. .G........ G..AT..... ..T.....T. .G..T..... ...C.....G
MAGNOLIA      ......T.A. .......A.. ...A.....A ..T.....T. ....A..... .........G
ORYZA         ..A..A..A. .G........ G..AT....A ........T. .A..A..C.. ...C.....A
MAIZE         ..T..A..A. .G........ GC..T....A ..T..T..T. .A..A..C.. C........T
ARABIDOPSIS   ..T..AT.A. .G...A.A.. .C.CT..... ..T..T.... ....G..... ......C..G

ZAMIA         CTGCCGGGTC AGCCAAGGGC AACTCAGCGA TCAGGACGCC CTATTAAGTC TATATGCAAT
CYCAD         .....T.... .......... .......... ..T.....G. .......... ..........
PINE          ..C..T.... .......... G......... ..G.....A. .C..C..... C.....T.G.
WELWITSCHIA   .....T.... .A........ .........G ..C..T..A. .A........ G..T....GC
PSILOTUM      .....A..C. .......A.. ...A..AA.G ..T.....A. .C........ .......C.G
NYMPHAEA      ..A..T.... ....T..... T..A..A... ..T...A.G. .......A.. ...T....G.
MAGNOLIA      ..T....... ....G..... T..A...A.G ........G. ....C..A.. C.......G.
ORYZA         ..T..T..C. .......... C..A...... ..T...A.G. .......A.. A..T....GC
MAIZE         ..T..T..C. .A..C..... C..A..A..T ..T...A.G. .......A.. A..T....GC
ARABIDOPSIS   T....T..A. .......... T......AA. .....GA.G. .......A.. A.....T.G.

ZAMIA         AGATTGAAGG CCAAAGAGGG TCGGATCAGA GGAAACCTGA TGGGTAAACG TGTAGATTTT
CYCAD         .......... ....G..... .......... .....TT... .......... ..........
PINE          ...C...... ....G..... ...C...C.. ..G..TT.A. ....G..... A.........
WELWITSCHIA   .......... .A..G..... .A........ .....T..A. ....A..... ...T.....C
PSILOTUM      C.T..A.... .A..G..A.. C......... ..G..T..C. ....G...A. G..T......
NYMPHAEA      ...C.T.... .T..G..... C.....T... ..G...T... ....G..G.. ..........
MAGNOLIA      C.GC.T.... .T.....A.. ......T... ..C...T... ....A..G.. ...G......
ORYZA         ...C....A. .A..G..A.. C..C..T... ......T... ....G..G.. ...T.....C
MAIZE         ..GC....A. .T.....A.. CA....T... ......T... ....C..G.. ...T.....C
ARABIDOPSIS   ..GC...... .A..G..A.. CA.A.....G ..T...T... ....A..... ...T.....C

ZAMIA         TCTGCACGTA CTGTTATTAC TCCAGATCCC AATATTAATA TTGATCAGCT TGGAGTACCC
CYCAD         .......... .......... ...T...... .....C.... .....G.... ......G...
PINE          .......... .A........ ...T.....A .....C..C. ........T. A.....G..G
WELWITSCHIA   .......... .......... C........A .....C.... .....G..T. G.....T..A
PSILOTUM      .....T..A. .......C.. A........T .....A..C. .....G.A.. ......T..A
NYMPHAEA      .....T.... ....C..... ...T.....A .C......C. .....G.A.. ...G..G..G
MAGNOLIA      ..A.....A. ....C..C.. A........G .CA.....C. .....G.A.. ......G..A
ORYZA         ..A..T.... ....C..C.. A.....C..G ........C. .....G.AT. G.....G..A
MAIZE         ..G..C.... .A.....C.. A........G ..C..A..C. .....G.AT. G........A
ARABIDOPSIS   ..G....... .......... .........A .CA..A.... .....G.A.. ...T..T..G
```

```
ZAMIA         TGGAGTATTG CCCTTAATTT AACCTATCCA GAAACTGTGA CTCCATATAA CATTGAAAGG
CYCAD         ........A. .......C.. G.....C..G .....A.... ....G..... ......G..A
PINE          .......... .AT.G..C.. G..A...... .......... .......... ...G..G..A
WELWITSCHIA   .....C..A. ....G..C.. G......... ....TC.... ....C..... ...A..G...
PSILOTUM      .......... .T..G..... ...A...... .......... .C........ T........A
NYMPHAEA      .......... .T......A. G..A.....T ........T. .......C.. ..........
MAGNOLIA      .......... .T........ ...A...... .......... .A........ ......G...
ORYZA         .......... .TT.G..CC. G..A..C... ........T. .......... ...C..G...
MAIZE         .......... .TT.G..CC. G..A.....T ........C. .......... T.....G...
ARABIDOPSIS   .......... .T..G...C. C..A..C... .....A..T. ....C..... .........A

ZAMIA         TTAAAGGAGC TTGTGGATAA TGGACCCCAT CCACCCCCAG GTAAAACAGG AGCCAGGTAC
CYCAD         .......... ....A..... ......T... ..T....... .A.....G.. T.........
PINE          .......... ....A..GT. ......T..C ..T..T.... .G.....T.. T....AA...
WELWITSCHIA   ........AT .A..T..AT. ...T..A... ..C..T.... .C........ T..A.AA..T
PSILOTUM      ..G..A.... ....T..GT. ......A..C ..T.....T. .C..G..... T..G..A..T
NYMPHAEA      ..G..A..A. ....T..G.. ......T..C .....G..T. .A.....T.. G..A.AA...
MAGNOLIA      ..G.....A. ....A..AT. ...G...... ..T..T..C. .......T.. T....A....
ORYZA         ..G..A..A. ....A..AT. ...G.....C ..T.....T. .G..G..... T..A.AA...
MAIZE         ..G.....A. ....T..AT. ...G..T..C ..T..Y.... .A..G..C.. T..A.AA..T
ARABIDOPSIS   .......... ....T...T. ......A... ..T..A..T. .G..G..T.. ...G.AA..T

ZAMIA         ATCATCAGGG ATGATGGTCA GAGGCTTGAT CTTCGTTACT TGAAAAAGTC TAGTGATCAC
CYCAD         ..T..T.... .......... ...A...... ..C..C...C .......... ..........
PINE          ..T....... .G.....C.. ...A.....C .....C.... ....G...AG C.....C..T
WELWITSCHIA   .......... .......... A..A...... .....C...C .A....GAAG C..C......
PSILOTUM      ..A..TC... .A.....G.. .C........ .......T.. .......AAG ......C.GT
NYMPHAEA      .....T.... .A..C..A.. .C.AG..... ..G..A..T. ...G....AG ..........
MAGNOLIA      .....A.... .A..C..A.. .......... .....G.... .......AAG .........T
ORYZA         ..T....... .A..C..... .......... ........TG ....G..AAG .........G
MAIZE         ..T....... .A........ .......... .....G..TG ....G..AAG C........G
ARABIDOPSIS   .....A..A. .......C.. A..A..A... .....G..TC .T..G...AG C........A

ZAMIA         CATTTGGAGT TGGGGTATAA GGTGGAGAGA CATTTGAATG ATGGAGACTT TGTTCTTTTT
CYCAD         .....A.... .......... ......A..G .......TG. .......... .........C
PINE          .....A.... ....T..C.. .........G .......... .......... .........C
WELWITSCHIA   ...C.A.... ....T..C.. .......... .....A..C. ....T..TC. ...C..G..C
PSILOTUM      ..C..A...C .T..T..... ......A..G ...C...... .......T.. ...GT.G..C
NYMPHAEA      ...C.T..A. .......... .......C.C .......T.. .......... ...C.....C
MAGNOLIA      ...C.....C .T..T..... .........G ..C....... ....G..... C........C
ORYZA         ...C.....C ....T..C.. ......A..G ..CC.C.... ....G..T.. .........C
MAIZE         ...C.....C .T..T..C.. .........G ...C.C.... ....G..T.. .........C
ARABIDOPSIS   ........AC .T..A..C.. .......C.G .....AC.G. ....T..T.. ......G...

ZAMIA         AATCGGCAAC CTAGTCTTCA TAAGATGTCT ATCATGGGGC ATAGAATCAA AATCATGCCA
CYCAD         .......... .......C.. .........G ........T. .........G ..........
PINE          .....A.... .......... C......... .......... .C..G..... ..........
WELWITSCHIA   ..C..A.... ....C..... C......... ..T.....A. .C..G..... ..........
PSILOTUM      ..C....... .......... ...A.....C ..T.....A. .C.AG..TCG T........T
NYMPHAEA      .......... .......A.. C..A...... ........T. ....G..... ...T.....T
MAGNOLIA      .......... .......... C..A...... ..A....... ..C....... ..........
ORYZA         .......... .C........ C..A...... .......... ..C.C..T.. .........C
MAIZE         .......... .A........ ...A...... .......... ..C.C..... ...T.....C
ARABIDOPSIS   .....T.... .A.....G.. C..A...... ........T. .C..G..T.G G..T......
```

```
ZAMIA         TACTCGACAT TTCGATTGAA CTTGTCAGTG ACTTCACCAT ACAATGCTGA TTTTGATGGA
CYCAD         .......... .......... .........C .....G.... .......... ..........
PINE          .....A..G. .....C.A.. .........C ..A..C.... .T........ .........G
WELWITSCHIA   ..T..A..G. .CA....... T.....C... .......... .T..C..... ..........
PSILOTUM      .T.AGT..C. .CA.GC.... TC.C..G..T ..AAGC..T. .......... ...C.....G
NYMPHAEA      .....T..C. .C..TC.A.. T..A.....T ..A....... .T........ .........G
MAGNOLIA      .....A..T. ....C..... T.....T..T .....T.... .......... .........G
ORYZA         ..T..A..T. .C..C..A.. ......C..C .....G..G. .......A.. .........G
MAIZE         .....A..T. .C..G..... ......T..C .......... .......... C..C..C..T
ARABIDOPSIS   ..T..C..T. .C..TC.... T.....T..C ..G..T..G. .......... .........G

ZAMIA         GATGAAATGA ACATGCATGT ACCCCAGTCA TTTGAGACAA GAGCAGAAGT CTTAGAATTG
CYCAD         .......... .......... .......... .......... .......... ...G...C.A
PINE          .......... .T.....C.. ...A...... ........G. .......... ...G...C.A
WELWITSCHIA   .....G.... .T........ G..T...... .....A.... .......... .A.G..GC..
PSILOTUM      .......... .T........ T..T...... .........C .G.....G.. T......C.A
NYMPHAEA      .......... .......... ...T....GT .....A..T. .......... ...G..G...
MAGNOLIA      .......... .......... T..T...... .....A..T. .......... ...G..GC..
ORYZA         .......... .T........ T..T...... ........C. ....T..... TC....G...
MAIZE         .....G.... .T........ C......... ........C. .......... TC.G..G..A
ARABIDOPSIS   .....G.... .......... ...A..A... ..C.....C. ....C..G.. G.....GC..

ZAMIA         ATGATGGTAC CAAAGTGCAT TGTCTCTCCC CAGTCGAATA GGCCTGTTAT GGGTATTGTC
CYCAD         ........T. .......... C........A .......... .A..G..... ..........
PINE          ........G. .......... C........T .....C.... ....A..... ..........
WELWITSCHIA   ........G. ....A..T.. ...T..C..A ..A..T..C. ....A..... .......C..
PSILOTUM      ........T. .C..A..... ...T.....T .....C...C .T..G..C.. T..G.....G
NYMPHAEA      .......... ....A..T.. ...T..A..A .....C...C ....G..C.. .........G
MAGNOLIA      .......... .G..A..... ......G..T ..A..A...C .C........ ..........
ORYZA         ........T. ....A..T.. ...G.....T ..A.....C. .......... ..........
MAIZE         ........G. ....A..... .........A ..A..A.... .......A.. ..........
ARABIDOPSIS   ........T. .T..A..T.. ......C... ...G.....C .T.....G.. ...A.....G

ZAMIA         CAAGACACTC TTCTAGGTTG CCGAAAGATC ACAAAGAGAG ACACATTTAT AGAGAAGGAT
CYCAD         .......... ..T.G..... .A........ .......... ....T..... ..........
PINE          ........A. ..T.G..C.. T..G.....A .....A.... .......C.. ...A..A...
WELWITSCHIA   .....T..CT .GT.G..C.. TA.....G.. .....AG... .T..T....C ..........
PSILOTUM      ........A. ....G..A.. TA.G..AG.T .....A.... .T..T..... ..........
NYMPHAEA      ..G..T..A. ....T..A.. TA.G...... ......C.T. ....C..C.. ..........
MAGNOLIA      ..G..T..A. .CT....A.. ...G.....A ..T..A.... .T..C..C.. T.........
ORYZA         ..G.....G. ....T..G.. T..C..A..T ..T..A.... .T..TC.... T..A......
MAIZE         ........A. .G..T..G.. T..C..A..T ..T..A..G. ....TC.A.. T..A......
ARABIDOPSIS   ..G..T..C. .CT.G..G.. ...T..A..T .......... .T..T..C.. ..........

ZAMIA         GTCTTCATGA ACATCTTAAT GTGGTGGGAG GATTTTGATG GCAAAATACC ATCYCCAACA
CYCAD         ..G....... .......... .........A A......... .......... ...T..G..T
PINE          .....T.... .T..T..... .......... .......... .A........ TG.T..G...
WELWITSCHIA   .......... .T..A..G.. .........A ..A.....A. .A..G..... TAAT...G.T
PSILOTUM      ..T..T.... ....TC.G.. .........A ..C....... .G..G..... TG.T.....T
NYMPHAEA      ..T....... ....TC.G.. .......... .......... .A..G..... TG.T..TG.C
MAGNOLIA      .....T.... .T.....G.. .......... .......... .G........ TG.T.....T
ORYZA         ..A..T.... .......... .........A .......... .A..GG.C.. TG.C..CG.C
MAIZE         ..A..T.... .......G.. .......C.A .....C.... .A..G..T.. TG.A..T..C
ARABIDOPSIS   ..A....... ...CAC.G.. .........A ..C..C.... .G...G.T.. GG.T..TG..
```

```
ZAMIA         ATCCTAAAAC CTAGACCTCT TTGGACTGGC AAGCAAGTAT TTAATCTTAT CATTCCAAGG
CYCAD         ..T.....G. .......... .......... .......... .C........ ..........
PINE          ..TTAC..G. .A..G..AA. .......... .......... .......G.. ........A.
WELWITSCHIA   ...T.G.... .A.....AA. .......... .......... .C.....G.. A.......AA
PSILOTUM      ..T..C.... .A.AG...T. G........G ......T... ...G...C.. T..A....A.
NYMPHAEA      ..TA.G..G. ........T. G.....A... ..A....... ....C..... A.......A.
MAGNOLIA      ..T..G..G. .......... ......A..A ........G. .C.....G.. ......T.A.
ORYZA         ..TT.G.... .A..G...A. .........G ..A.....T. .C..CT.A.. T.....C.A.
MAIZE         ..TT.G.... ....G...A. .........G ..A.....T. ....CT.A.. T..C..C.A.
ARABIDOPSIS   ...T.G..G. ..C.T..... .......... ..A.....T. .......... ...A....AA

ZAMIA         CAGATAAATC TCATAAGATA TTCTGCATGG CATTCRGAGT CTGAGACAGG ATTTATTACA
CYCAD         ........C. .T........ C......... .....A.... ....A..... ......C..G
PINE          .....T.... .T.....GAC ...A...... .....A.... ....AT.G.. .G........
WELWITSCHIA   .....T.... .T........ .......... ..CAAT..AA GC....A... ....CC...T
PSILOTUM      .....C.... .A...C.T.T ...A..C... ....TG..T. .....GTT.. TGAG.....T
NYMPHAEA      .........T .A..T..G.. C..A...... .....G..A. .A..A..T.. ...C.....C
MAGNOLIA      ........C. ........AC C..G...... ..C..G..AG .G..A..... ......C..T
ORYZA         ..A..C...T .G..T....T .....GC... .....T..AG ....A...A. .........T
MAIZE         ..A......T .A..TC.G.T ...A...... .....G..AG AA..A.A... .........T
ARABIDOPSIS   .......... .GT.G..G.. C.....T... ..CG.A..TA .A.....T.. ......A..T

ZAMIA         CCAGGGGATA CCGTTGTTCG TATTGAGAAA GGAGAAGTTC TTTCAGGCAC TCTCTGTAAA
CYCAD         .......... .TTG...... G.....A..G .......... .G..G..... A.........
PINE          .....A.... .TC....A.. C........G .........A .C........ ...T..C..G
WELWITSCHIA   ..T..A..C. .T.....C.. G......... ........G. ....T..T.. ...T..C..G
PSILOTUM      .......... .T.....CA. A.....A..G .....GC.CA .......... C..T..C..G
NYMPHAEA      ..T..A.... .TTG...C.. A..A....GG ......C... .C.....G.. C........G
MAGNOLIA      .....A.... .ACAA...A. A..A....G. ..C..GC.G. ..G.T..... ......C..G
ORYZA         ..T..T.... .TA.G..C.. G..A.....G .....GC... .A..T..T.. A..T..C...
MAIZE         ..C..T.... .TA.G..CA. G..A.....G .....GC... .G..T..... A..T..C...
ARABIDOPSIS   ..G....... .TCAA..G.. A.....A.G. ..G...C... ..G.C..A.. ...T..C...

ZAMIA         AAAACACTCG GAACATCTTC TGGAAGTCTT ATTCACGTGA TCTGGGAGGA GGTTGGTCCA
CYCAD         .....C..T. ....G..... .......... .....T.... .......... ..........
PINE          .....C..T. ....T..... A......... .....T.... .......... ...G......
WELWITSCHIA   ........T. ..T.T..... A......... .....T.... .......A.. ..........
PSILOTUM      ..G..T..T. .T..T...GG ...C...T.G .....T.... .T.....A.. A..G..G..T
NYMPHAEA      .....CA.G. .C.....A.. .......... ........T. .......A.. ...G......
MAGNOLIA      ..G..C..T. ........A. ...T...... .....T..C. .......A.. ..........
ORYZA         ..G.....T. .......AA. .......... .....T..T. .T.....A.. ..........
MAIZE         ..G.GT..T. .....GGC.. C......... ..C..T..C. .T.....A.. ..........
ARABIDOPSIS   ..G..C..T. .T......AA .........C G.G..T..C. .T.....A.. .........T

ZAMIA         GATGCAGCTC GCAAGTTTTT GGGGCACACA CAATGGCTTG TTAACTATTG GCTTTTGCAG
CYCAD         .......... .......... ......T... ..G....... .......... ...A.....A
PINE          .....T.... .T........ A..T...... .......... .......C.. ...GC.A...
WELWITSCHIA   ........A. .A..A..... A..T.....C ..G....... .A........ ...G......
PSILOTUM      ........G. .T..A...C. ...T...... ..G.....A. .A........ .T.G......
NYMPHAEA      .....T..A. .......... ......T..T ..G....... ....T..... ....C.....
MAGNOLIA      .....T..C. ....A..C.. ...C...... ..G....... .......C.. ..........
ORYZA         .....T.... .......C.. A..T..T... ..G.....A. .C.....C.. ....C....A
MAIZE         .....T..C. .G.....C.. A..A...... ..G....... .A.....C.. ....C.T..A
ARABIDOPSIS   .........A .A..A..CC. C..T..T..T .......... .C..T..C.. ....C.....
```

```
ZAMIA         CAAGGTTTCA GTATTGGTAT AGGAGACACC ATTGCTGATG CTGCAACAAT GGAAACAATT
CYCAD         ..G....... .......... .......... .......... .......... ..........
PINE          ..G....... .......A.. .........A .......... .......... ...GGTT..C
WELWITSCHIA   ..G....... ....A..C.. T.....T..T .....A.... .......C.. .....A....
PSILOTUM      ..T....... .C.....C.. T..T..T..A .....A..C. .......T.. ...G.AG..C
NYMPHAEA      A.T.C...T. .C........ T..G...... .....A.... ..T.T..C.. .....AG...
MAGNOLIA      A.T..A..T. .......A.. T..G.....A .....A.... .AT....T.. .....A....
ORYZA         A.T....... .......A.. T..G..T..A .....A.... .A..C..C.. ...G.AT...
MAIZE         A.T....... .......A.. T..G..T..T .....A.... .AT.C..C.. ..........
ARABIDOPSIS   A.T.....T. CC..C..A.. T..T.....A .....C...T .AT....... ...G.A....

ZAMIA         AATGAAACAA TTTCGAAAGC AAAGAATGAA GTGAATCACC TTATTCAGCT TGCTCATCAA
CYCAD         .......... .C..A..... ....GC.... ..C.....A. .......... ......C..G
PINE          .......... .A..A..... ...A...... ..C..AATG. ....CA..GC ......GG.G
WELWITSCHIA   ..C.....T. .AG.A..T.. T..A.T.A.. ..CC.G..G. ....AG..AA A..C..AG.G
PSILOTUM      .......GT. ..G.A..... ...A.....G ..C..AG.G. .G...A.AA. G..A..AG.G
NYMPHAEA      ..C..G..G. .C..T..... ...A.....G .....AG.G. ....CA..GC ...C..GG.G
MAGNOLIA      .......... .A........ G......... .....GG.A. .....A..GC ...C..AG..
ORYZA         .....G.... .C..A..... T..A.....T .....GA.A. .....A...A ATT..G.G.T
MAIZE         .....T.... .A..T..... T.......CT .....GG.G. .....A.AAA A......G.G
ARABIDOPSIS   ........T. ....C..T.. ...A.C..CT .....AG.T. ....C.G..A GTTC..GGG.

ZAMIA         AAAGCATTAG AGGCAKAGCC TGGGCGTACG ATGATGGAAT CTTTTGAAAA CAGAGTCAAT
CYCAD         .......... .....G.A.. ......C... .......... .A........ ..........
PINE          ...CA..... .....G.A.. ...C..C..A ........G. .......G.. ..........
WELWITSCHIA   ...AAGC.G. .A...G.A.. ...A.....T ........G. .A........ .CA......C
PSILOTUM      ..GCA..... .....G.A.. A..A..A..A .C........ .......G.. TC....T...
NYMPHAEA      ...CAG..G. .A..TG.A.. ......A..C ........G. .A.....G.. T.........
MAGNOLIA      ..GCAGC... .....G.A.. ......A..T .......... .A.....G.. T..G..G..C
ORYZA         ..CCA...G. .A..TG..G. A..A..C..C .C........ .A........ T.....A...
MAIZE         ..GCAG..G. .A..TG.... A..A..C..T .......... .A........ ......G..C
ARABIDOPSIS   ..G.A...G. .CC.TG.... ...C..A..T ....GA..TA .A.....G.. ...G..T..C

ZAMIA         CAGGTGTTGA ATAAGGCTCG TGATGATGCA GGAAGTAGTG CCCAAAAGAG CTTATCAGAA
CYCAD         .......... .......C.. .......... .......... .T.....A.. .........G
PINE          .......... .C.....C.. .......... .......... .T.....A.. T........G
WELWITSCHIA   .......... .C.....C.. .........G ..T.A...C. .T..G..A.. .........G
PSILOTUM      .......... .C.G...... .......... .....C.... .G..G.G... ...G..T..G
NYMPHAEA      .......... .C..A..... ......C..T ..T....... .T..G.G... ......T...
MAGNOLIA      ......C... .......... .........T ..G....... .A..G..... ..........
ORYZA         G....TC.T. .C........ .....T...T .......... ..G.G..... T..G..T...
MAIZE         .....TC.T. .C..A..C.. .........T ..T..C.... .T..G..T.. ...G..T...
ARABIDOPSIS   .....T.... ....A..... .........T .......... .T........ T...G.....

ZAMIA         AGCAATAATT TGAAGGCAAT GGTCACTGCT GGCTCAAAAG GAAGCTTTAT TAATATATCA
CYCAD         ..T....... .......... ...T...... .......... ....T..... ..........
PINE          ..T....... .......... ...T..A..C ..G....... ....T..... ...C......
WELWITSCHIA   ..T......C .......... ...T..A..A ..G..T.... ....T..... C.....T..C
PSILOTUM      ..T......G .C........ ...T..A... ..A.....G. .CTCT..C.. C..C..C...
NYMPHAEA      ..T.....CC .......T.. .........A ..A..T.... ....T..... ......T...
MAGNOLIA      ..T..C...C .......T.. ...G..A..A ..A.....G. ....T..... C..C..T...
ORYZA         ..T..C...C .......T.. ..C...C..A .......... ...CT..C.. C.....T...
MAIZE         .....C.... .......T.. .........A ..T....... .C..T..C.. ...C..T...
ARABIDOPSIS   .C......CC .T.....C.. ...G..A..A ..A..C.... ....T..C.. C.....T..T
```

```
ZAMIA        CAAATGACTG CTTGTGTGGG ACAGCAAAAT GTTGAAGGGA AACGGATTCC ATTTGGACTT
CYCAD        ..G....... .......... .......... ........A. .......... ......TT..
PINE         .......... .......... ...A...... ..A.....A. ....A..... T.A....T..
WELWITSCHIA  ..G....... .......... G......... ........T. .......... T.A...CT.C
PSILOTUM     .......T.. .A.....T.. T..A...... ..G.....C. .G..C..A.. ......GT..
NYMPHAEA     ..G....... .......... T..A..G... ..G..G.... .G..A..... T.....GT.C
MAGNOLIA     ..G....... .......... ...A..G... ..C..G..T. ....A..... T.AC..TT.C
ORYZA        ..G....... .......C.. ......G... .....G..C. .G.....C.. ......CT.C
MAIZE        .......... .......... ......G... .....G..C. .G.....C.. T.....TT.C
ARABIDOPSIS  ........A. .G.....C.. T......... .......... .G..A..... .......T..

ZAMIA        TTTGGGCGAT CCCTACCCCA TTTCACAAGG GATGATTATG GTCCTGAAAG TCGTGGTTTT
CYCAD        ..C....... .......... ......G... .....C.... ....A..... ..........
PINE         GAC..T...A .T..T..... ...T....A. ......A.A. ....A..... ..........
WELWITSCHIA  AA....A.GA .A..G..T.. ........A. .....CA.G. ....A..G.. C.........
PSILOTUM     G...AT..GA .A..G..A.. ......G.A. .....C.... .A......TC ..........
NYMPHAEA     A...AC..CA .A..G..A.. ...T..T.AA .......... .G........ ..........
MAGNOLIA     A...AT..GA .G.....A.. ......C.AA .....C.... .G..C..... ...G..C...
ORYZA        ACCAAT...A .AT.G..... C.....G.AA A....C.... .......... ......G...
MAIZE        A...AT..GA .GT.G..A.. ........AA .....C.... .......... ......A...
ARABIDOPSIS  GA......GA .AT.G..A.. ......C.AA .......... .......... ..........

ZAMIA        GTGGAAAATT CTTATCTCCG TGGCCTAACA CCGCAAGAGT TCTTTTTCCA CGCAATGGGT
CYCAD        .....G.... .......G.. ...T...... .......... .......... T........A
PINE         .....G..C. .......T.. ......G..T ..C..G.... .......T.. .........A
WELWITSCHIA  .....G..C. .......T.. ...AT.G..T ..T..G.... .......... T..T.....A
PSILOTUM     .....G.... .C...T.GA. A..TT....T ..T....... .T..C..T.. T..T.....A
NYMPHAEA     .....G..C. .C.....T.. ...G..G..T ..T..G..A. .T.....T.. T..T......
MAGNOLIA     .....G..C. .A........ ...G..G..T ..A..G..A. ....C..T.. ...T......
ORYZA        .....G..C. ....C..T.. A.....G... ..A.....A. .T..C..T.. T..T......
MAIZE        .......... ....C..... A..T..G... ........A. .T..C..T.. ...T......
ARABIDOPSIS  ..T..G.... .G..C..G.. ....T.G..T ..T....... .......... T..T.....A

ZAMIA        GGTCGAGAAG GTCTGATAGA TACTGCTGTG AAAACTTCAG AAACAGGGTA TATTCAGAGG
CYCAD        ..A....... .......... .......... ........T. .G........ C..A......
PINE         ...A...... .GT....... C......... ..G.....G. .......T.. C..A......
WELWITSCHIA  ...A...... ..T.A..... .......... ..G....... .......T.. C..C..A..A
PSILOTUM     ...A...... ....A..T.. ...G...... ........T. .G..T..T.. ...A.....A
NYMPHAEA     ...A.G.... ..T....T.. ......A..C .......... .G..T..T.. C..A...C..
MAGNOLIA     ...A.G.... .......T.. C.....A... .....A..T. .G..G..... ..........
ORYZA        ...A...... .......T.. .......... ..G..C..G. ....T..A.. ...C...C.A
MAIZE        ...A.G.... ....T..... C......... .....C..T. .G..T..... ...C......
ARABIDOPSIS  ..A....... ....T..T.. .......... ..G..A.... ....T..A.. C.........

ZAMIA        CGTCTTGTAA AGGCTATGGA GGACATCATG GTTAAATACG ATGGGACGGT CAGGAACTCA
CYCAD        ........G. .......... .......... ........T. .......A.. ......T...
PINE         .......... ....A..... ...T...... ........T. .C.....T.. ..........
WELWITSCHIA  .......... .......... ...T..T... ........T. ....C..T.. G..A.....C
PSILOTUM     A.GT.G..T. ....A..... ....G.T... ..G..G..T. ....T..A.. .C.C..T...
NYMPHAEA     .......... .A........ .......... ..C.....T. ....T..A.. T........G
MAGNOLIA     .....C..G. .......... A.....T... ..C.....T. ....T..T.. T.....T...
ORYZA        ..A.....G. .......... A..T...... ..G..G..T. ....T..T.. T..A..T..T
MAIZE        A.A.....G. .A........ .......... ........T. ....T..T.. A..A..T..T
ARABIDOPSIS  ..AT.G.... .......... ...T..T... .....G..T. .......A.. ...A.....T
```

```
ZAMIA        TTAGGCGATG TTATTCAATT TCTCTACGGG GAAGATGGTA TGGATGCTGT TTGGATTGAG
CYCAD        .....T.... .......G.. ...G..T... .......... .......... ..........
PINE         C.T..T.... .......... ...A..T..T .......... .......... .........A
WELWITSCHIA  C.T..T.... ...CC..G.. .T.G..T..A .....C..C. .......... C........A
PSILOTUM     ..G..G.... .....G.G.. C.....T..T .......... .......A.. ..........
NYMPHAEA     .....A.... .G..C..G.. ...T..T... ..G....... .......... .........A
MAGNOLIA     .....G.... .G.....G.. .T.G..T..A ..G....... .......... .........A
ORYZA        ..G..A.... .C.....G.. CT.A..T..A ........C. ........A. ..........
MAIZE        C.T..A.... .C..C..G.. CT.G..T..A ........C. .......... .........A
ARABIDOPSIS  ..G..T.... .......... ......T..A .......... .......... A.....A..A

ZAMIA        AGTCAGAAAC TTGATTCATT GAAGATGAAA AGGCCAGAGT TTGAAAATGT TTATAGGTAT
CYCAD        .....A..G. .C........ .......... ...T...... .......... ..........
PINE         TCA..A..G. .......TC. T......... .A.AAG.... .......... .....AA...
WELWITSCHIA  TCA.....G. .......T.. A......... .A.AG...A. ....TG.... ...C.A...C
PSILOTUM     TCC.....G. .G......A. ...A.....G .CAAA..... ....C.GAA. G..C.T..T.
NYMPHAEA     TCG....... .A........ ...A...... .A.T.T.... ....T..G.. C.TC.A..T.
MAGNOLIA     TCA.....GT .G.....TC. ...A...... .A..GT.... .C..T..... ..TC......
ORYZA        TCA.....GT .G..C..C.. .......... .A.G.T.... ....T..... A.T.C.T...
MAIZE        TC...A...T .G..C..... .........G .A...T.... .C..T..... A.TCC.T...
ARABIDOPSIS  TCA.....G. .G.....C.. ...A.....G .AAT...... ....T.GGAC ..T..A....

ZAMIA        GAGATTGATA ATGAAAACTG GAATCCCAGC TACATGTTAC CAGAGCATGT TGAAGATTTG
CYCAD        .......... .......... .........T .......... ....A..... ..........
PINE         ...C.....C .A.....T.. ......T..T .......... ....A..... ..........
WELWITSCHIA  ...C.....C .G..G..T.. ......TG.T ..T.....G. .T..A..CA. ......C...
PSILOTUM     .........T C...G..... ......AGAT ........GG .T........ G..G..C...
NYMPHAEA     ...T.....G .......T.. .G....A..T ..TT..CA.. .T..TTT... G.....C..A
MAGNOLIA     ..AT....C. ...CC..T.. ......A..T .......... .T........ ...G......
ORYZA        ..AC.G...G ....G..... ...G..A.A. ...T...CGA .TC.A....C ...G......
MAIZE        ...C.G...G ....G..T.. ..GA..T.A. ......C... .T..A..... ...C......
ARABIDOPSIS  ........CG .C........ ......T.CT ...C.AAGTG AT..A...C. ......C...

ZAMIA        AAAACCATCC GAGAATTCCA GAATGTGTTT GATGCAGAAG TGCAGAAACT GGAAGCAGAT
CYCAD        ........T. ....G..T.. .......... ......C... .......G.. ...G..T...
PINE         .....A...A .G..G..... ......A... .......... .......... ...GT.....
WELWITSCHIA  ..G..A..T. ....G..... ......T.G. .......... ....AC.T.. ....AAT..C
PSILOTUM     .....T..T. AG..G....G T......... ........G. .T........ A..G.....C
NYMPHAEA     ........TA A...GC.T.. .G.......C C.C....... .T.....G.. C...A.....
MAGNOLIA     .......... .G..G....G C.....T... .....T.... .T........ ..........
ORYZA        ..G......A .T..GA.TAG A......... ..G.....G. .T.....G.. A.....T..C
MAIZE        ..G.....A. .T......AG A........C ..G.....G. .T..A..G.. A.....T...
ARABIDOPSIS  ..GGGG..T. .G..G..G.G TG....A..C .....G...T ATTC...... T..GA.T..C

ZAMIA        AGACGGCAGC TAGGAACAGA GATTGCTCCT AGTGGAGATA ACTCTTGGCC TATGCCTGTG
CYCAD        .......... .......... .......... .......... .......... C.........
PINE         .....A.... ....T.TG.. ......A... .....G.... ....G..... .C.......A
WELWITSCHIA  ..G..T..A. .......T.. A..A.....A .....G.... .T........ CT........
PSILOTUM     ........A. .GT.C..... ......C... TCA..T..C. .T..A..... AT....A..T
NYMPHAEA     ..G....... .T.CC..... .......A.C .C...G..C. .T..A..... A.....G...
MAGNOLIA     C..TAC..A. .T..T..... .......A.. .CA..T.... ....A..... .........T
ORYZA        C.GTTC.... .T..C..T.. .......A.A .C...T.... .TA.A..... .........C
MAIZE        C.GTAC.... .T..GT.T.. ...CA.CA.A .C...T..C. .T..G..... ..........
ARABIDOPSIS  ...TTC..A. .C..G..... A.....AA.A .A...T.... G.A....... AT.......T
```

```
ZAMIA         AATCTGAAGA GGTTAATATG GAATGCACAG AAAACATTCA AAATTGATTT GAGAAAGCCA
CYCAD         .......... ..C....C.. ......C... .......... .......... .........T
PINE          ..C..T.... ..C.T..... ....AGC... ..G..C.... ........GC .....GA..C
WELWITSCHIA   .....C.... ..C.T..T.. ......T... ...TTG..T. ........CC A.....AG..
PSILOTUM      .....T.... .A..G..T.. ......T... ..G.TT.... .GG.G..... C.....A..T
NYMPHAEA      .....C..A. ..C.T..C.. .........A ..G..G..T. .GG....CC. .....GA...
MAGNOLIA      ..C..C.... ..C.T..C.. ......T..A ..G.....T. .GG....C.. .....G....
ORYZA         .....C...C ..C.T..C.. ......T... ..G....... .G......C. .....G...T
MAIZE         ..C..C...C .AC.C..C.. ......T... ..G....... .G........ T....GA..T
ARABIDOPSIS   ..CA.C.... ..CAT..C.. ......G... ..G..T.... .......C.. .C.C..AATT

ZAMIA         TCAGATATGC ATCCCATGGA GATTGTGGAA GCTGTTGACA AGCTTCAGGA AAGGCTTAAA
CYCAD         .......... .......... .......... .......... .......... ..........
PINE          .....C.... .......... ......A... ..AA...... GAT.G..... ...A.....G
WELWITSCHIA   .......... .C........ A...A.T..T ..CA....T. .......... ...A.....G
PSILOTUM      .......... .C........ ......A... ..A.....T. ....C..... ...AT.G...
NYMPHAEA      ..T....... .C..T..... ...A...... .......... .A........ ......G...
MAGNOLIA      ..T..C.... .C..T..... ......T... ..A....... .......... ....T.G...
ORYZA         ..T..C.... .C..A..... A........T ..AA.A..T. ....G..A.. ...A.....G
MAIZE         ..T..C.... .C..G..... A......... ..GA.A..T. .A..G..A.. ...A.....G
ARABIDOPSIS   .......... .C..TG.T.. A.....T..T ........T. .A..A..... G.....GTTG

ZAMIA         GTGGTTCCTG GTGATGACCT TATGAGTATA GAGGCCCAGA AGAATGCTAC TCTCTTCTTT
CYCAD         ..T....... .C........ .......... ........A. .......... ......T...
PINE          ..T....... .......... C........T .....A.... .A.....C.. ......T..C
WELWITSCHIA   .....C..A. .CA......C AT........ .......... .......... .........C
PSILOTUM      ..T....... .A..G..TGA A..T..C.AG ..A.....A. .......... A..G..T..C
NYMPHAEA      ..T..A.... ....G..TGC A........G .....T..A. .......... C..T..T..C
MAGNOLIA      ..C.....A. .......TT. .C....C..G ..A..T.... .A.....A.. C..T.....C
ORYZA         ..T..C.... ....C..TGA ...C.....T .....T.... .......A.. .T.G.....C
MAIZE         ..T..A.... .......TGC ......C..T .......... .......A.. CT.G......
ARABIDOPSIS   ..T....... .......TGC GT.....G.G ..A..A.... .A..C..A.. AT.G......

ZAMIA         AACATTCTAC TTCGCAGTAC ATTTGCCAGT AAACGTGTAT TAAAGGAATA CCGGCTTACC
CYCAD         ......T... .......... ......T... .......... .......... ..........
PINE          ..T...T.G. ....A..... ......T... .........C .G.....G.. ..AC..C..A
WELWITSCHIA   .....CT... ....T..... .......... .....G..G. .G........ ..AC..G..T
PSILOTUM      .....CT.GT .G........ CC.G...... ........TC .G........ .A....C...
NYMPHAEA      ..T..A..C. ..A.A..C.. ......A..C ..G.....G. C...A..G.. TA.......A
MAGNOLIA      ........G. ....T..C.. .......... ..G.....G. .G..A..G.. .A....A..T
ORYZA         ..T..C..G. ....A..C.. ......T..C ..GA.G..C. .G.....G.. .A.......A
MAIZE         ..T..C.... ....T..C.. G.....T..C ..GA.G..C. .G........ .A.......A
ARABIDOPSIS   ......T.G. .......C.. TC....T... ...A.A..G. .GG.A..... .AA...C.G.

ZAMIA         AGAGAAGCAT TTGAATGGGT GATTGGTGAG ATTGAGTCTC GTTTCCTGCA GTCTCTCGTG
CYCAD         .......... .......... .......... .......... .......... A...T.G...
PINE          .......... .......... AG........ .....A.... .......... ...CT.A..A
WELWITSCHIA   .A........ ....T..... A........A .....AA.G. .C..T..A.. ....T.A..A
PSILOTUM      .A........ ....G..... T..A..G... .....AA.C. .....T.A.. ...AA.T..T
NYMPHAEA      C.T....... ....C..... .......... .......... .C..T..... ...A..T..T
MAGNOLIA      C.T....... .......... A..C...... ..A..A.... .A...T.A.. A..AT.G..T
ORYZA         .AG.....T. .......... T......... .....A..GA .A.....T.. A...T.A...
MAIZE         .AG.....T. .......... T......... .....A..CA .G..T..T.. ....T.G...
ARABIDOPSIS   C.C..G..T. ....G..... C......... .....A..AA .G..TT.A.. A..G..A...
```

```
ZAMIA        GCACCAGGTG AGATGATTGG TTGTGTAGCA GCGCAGTCCA TT
CYCAD        ..T....... .......... ...C...... ..A....... ..
PINE         ..T....... .......... C.....T... ..A....... ..
WELWITSCHIA  T....G..A. .......C.. ...C..T... ..T....... ..
PSILOTUM     ..T..T..G. .A........ C.....T..T ..T..A..T. ..
NYMPHAEA     ...G...... .......... C..C...... ..A..A..G. .A
MAGNOLIA     .......... .A........ C.....G..T ..C..A.... .A
ORYZA        ..T....... .A.....C.. A........T ..A....... .C
MAIZE        ..C..T..C. .......... A.....G..T ..A..A..T. ..
ARABIDOPSIS  ..C.....G. .A.....C.. ......T..T ..T..A..A. ..
```

Figure 3 The RPB1 nucleotide sequence alignment of the ten plant species used in this study. All nucleotides identical to those in the *Zamia* sequence are represented by a dot and all dashes correspond to gaps in the alignment.

show how easily the protein coding sequence of RPB1 can be aligned at both the nucleotide and amino acid levels, due to the presence of very few gaps, even when distantly related species are compared. There are only five small indels that appear in the first 125 amino acids or 375 base pairs. The sequences of the outgroup, *Psilotum*, all the gymnosperms and *Nymphaea* are each 3039 base pairs long, the *Arabidopsis* sequence is 3036 base pairs long, the *Magnolia* and *Oryza* sequences are 3030 base pairs long and maize has the shortest sequence at 3027 base pairs (Table 1, Fig. 3).

Sequence analyses:

When analyzing the G + C content of the ten plant taxa, it is clear that there is no G + C bias at any codon position compared to the total G + C content of the gene (Fig. 4). The total G + C content of the RPB1 gene among all plants varies from 42% to 43%. The first codon position shows the least variability in G + C content ranging from 51% to 52%, the G + C content at the second codon position varies from 36% to 38% and the third codon position shows the largest variability in G + C content ranging from 38% to 43%. The gymnosperm plant species and the outgroup *Psilotum* contain between 38% and 40 % G + C at the third codon position and the angiosperms, excluding *Arabidopsis*, have between 41% and 42% G + C at the third codon position (Fig. 5). *Arabidopsis*, with a value of 39%, has a slightly lower G + C content at the third codon position compared to the other angiosperms studied (Fig. 5).

Figure 4 The frequency of G + C at each codon position compared to the total G + C frequency in the A to G region of RPB1. The slope (m) and R^2 value are shown for each codon position.

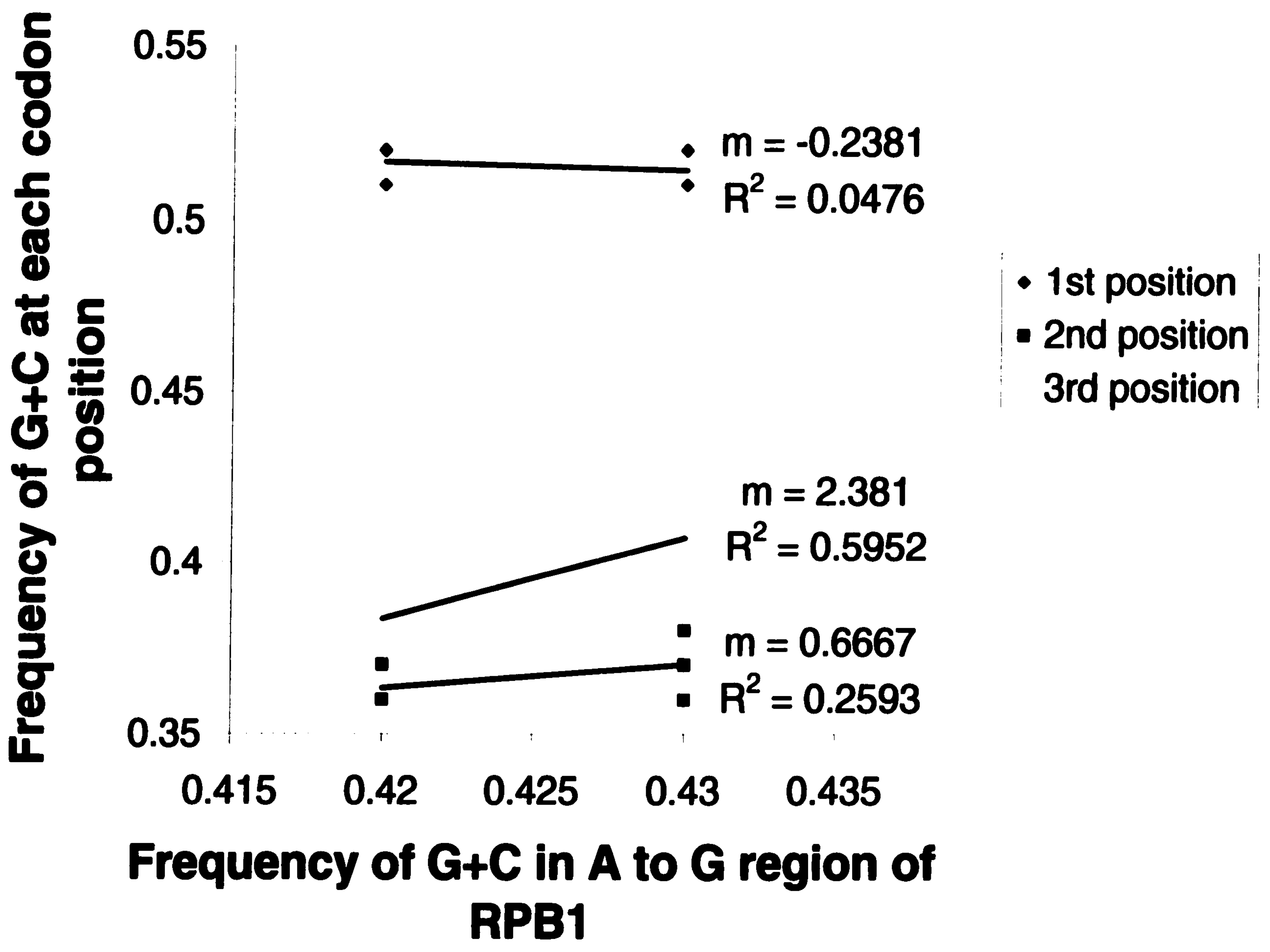
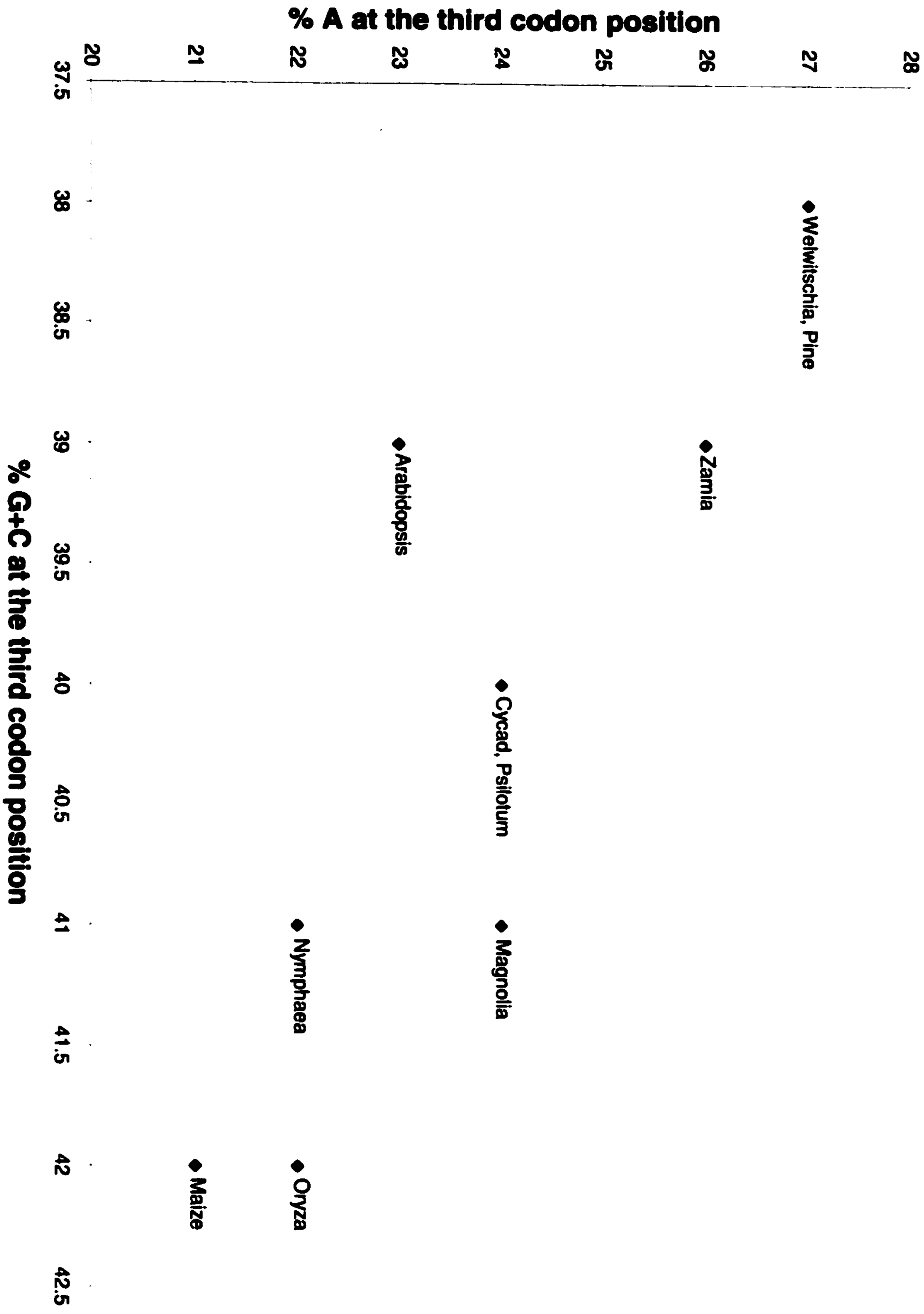

Frequency of G+C at each codon position
0.55
0.5
0.45
0.4
0.35
0.415
0.42
0.425
0.43
0.435
Frequency of G+C in A to G region of RPB1
m = -0.2381
R^2 = 0.0476
m = 2.381
R^2 = 0.5952
m = 0.6667
R^2 = 0.2593
◆ 1st position
■ 2nd position
3rd position

Figure 5 The percent of G + C at the third codon position compared to the percent of A at the third codon position for RPB1 in all ten plant species.


% A at the third codon position
20
21
22
23
24
25
26
27
28
% G+C at the third codon position
37.5
38
38.5
39
39.5
40
40.5
41
41.5
42
42.5
Welwitschia, Pine
Zamia
Arabidopsis
Cycad, Psilotum
Magnolia
Nymphaea
Oryza
Maize

Synonymous (Ks) and nonsynonymous (Ka) substitutions per nucleotide site were calculated for RPB1 in all ten taxa (Table 2). *Arabidopsis* has accumulated the largest number of synonymous substitutions per nucleotide site, as the value was too high to be calculated when compared to the gymnosperm and outgroup species. All comparisons contained more than one synonymous substitution per nucleotide site except for *Zamia* / *Cycad*, *Zamia* / pine, *Cycad* / pine and *Oryza* / maize. The nonsynonymous substitutions per nucleotide site for RPB1 varied from 0.0115 ± 0.0023, when *Zamia* and *Cycad* were compared, to 0.1092 ± 0.0073, when *Psilotum* and *Arabidopsis* were compared.

Relative rate tests were performed to determine whether any RPB1 sequences were evolving at significantly different rates than other RPB1 sequences. With the α value adjusted for multiple comparisons by the Bonferroni correction (α = 0.05 / 36 pairwise comparisons = 0.0014), *Arabidopsis* and *Oryza* reveal significantly different evolutionary rates, but all other plant RPB1 sequences evolve at relatively the same evolutionary rate (Table 3). When all gymnosperm species and all angiosperm species are combined into two separate lineages, the relative rate of the RPB1 sequences is found to be statistically significant with a probability value equal to 0.0042 ± 0.0075 (Table 4). However, when *Arabidopsis* is removed from the angiosperm lineage, the probability value is equal to 0.1411 ± 0.0076 and therefore, is no longer statistically significant (Table 4).

Table 2. Estimated synonymous (Ks) and nonsynonymous (Ka) substitutions per nucleotide site for all RPB1 sequences. The values are displayed ± their standard deviation. All values that were too high to be calculated are represented by an asterix (*).

Ks

	Zamia	Cycad	Pine	Welwitschia	Psilotum	Nymphaea	Magnolia	Oryza	Maize	Arabidopsis
Zamia	0	0.3146± 0.0264	0.7749± 0.0550	1.1951± 0.0926	2.7248± 1.4218	1.5537± 0.1725	1.3866± 0.1301	2.1627± 0.3872	3.0258± 2.8343	*
Cycad	0.0115± 0.0023	0	0.7178± 0.0517	1.0865± 0.0794	2.2475± 0.5340	1.6562± 0.2360	1.4726± 0.1344	1.8251± 0.2171	2.1991± 0.3699	*
Pine	0.0418± 0.0044	0.0420± 0.0044	0	1.0911± 0.0822	2.0413± 0.3508	2.0193± 0.4625	1.6600± 0.1947	2.1802± 0.4416	*	*
Welwitschia	0.0541± 0.0050	0.0564± 0.0051	0.0418± 0.0044	0	*	2.4305± 0.7483	1.8890± 0.2119	2.2810± 0.4452	2.7714± 1.1923	*
Psilotum	0.0696± 0.0057	0.0695± 0.0057	0.0758± 0.0060	0.0845± 0.0064	0	2.3945± 0.6076	2.4987± 0.7370	*	*	*
Nymphaea	0.0522± 0.0049	0.0514± 0.0049	0.0537± 0.0050	0.0672± 0.0056	0.0726± 0.0059	0	1.1789± 0.0952	1.6393± 0.1834	1.7680± 0.1845	2.5155± 0.7751
Magnolia	0.0461± 0.0046	0.0485± 0.0047	0.0477± 0.0047	0.0652± 0.0056	0.0714± 0.0058	0.0410± 0.0043	0	1.2135± 0.0890	1.4401± 0.1201	1.8786± 0.2381
Oryza	0.0721± 0.0058	0.0731± 0.0059	0.0708± 0.0058	0.0834± 0.0063	0.0921± 0.0067	0.0657± 0.0055	0.0528± 0.0049	0	0.5317± 0.0374	2.0198± 0.2777
Maize	0.0611± 0.0053	0.0623± 0.0054	0.0613± 0.0053	0.0744± 0.0060	0.0814± 0.0062	0.0567± 0.0051	0.0430± 0.0044	0.0316± 0.0038	0	2.9152± 1.6465
Arabidopsis	0.0897± 0.0066	0.0890± 0.0066	0.0878± 0.0065	0.0963± 0.0068	0.1092± 0.0073	0.0825± 0.0063	0.0699± 0.0058	0.0885± 0.0066	0.0872± 0.0065	0

Ka

Table 3: Relative rate probabilities ± standard deviation of all pairwise comparisons for the RPB1 amino acid sequences present in this study. *Psilotum* was the outgroup in all rate calculations. Substitution rates are significant at $P < 0.0014$.

Species/ Species	Zamia	Cycad	Pine	Welwitschia	Arabidopsis	Oryza	Maize	Nymphaea	Magnolia
Zamia		0.8179± 0.0049	0.5389± 0.0092	0.0189± 0.0107	*5 E^{-7}± 0.0135	*0.0004 ±0.0124	0.0849± 0.0113	0.3997± 0.0108	0.8191± 0.0099
Cycad	0.8179± 0.0049		0.4546± 0.0091	0.0147± 0.0108	*7 E^{-7}± 0.0132	*0.0003 ±0.0124	0.0695± 0.0113	0.3385± 0.0107	0.7317± 0.0099
Pine	0.5389± 0.0092	0.4546± 0.0091		0.0436± 0.0097	*2 E^{-6}± 0.0133	0.0016± 0.0121	0.2048± 0.0108	0.7493± 0.0107	0.7328± 0.0100
Welwitschia	0.0189± 0.0107	0.0147± 0.0108	0.0436± 0.0097		0.0017± 0.0140	0.1622± 0.0134	0.6458± 0.0126	0.1867± 0.0122	0.0514± 0.0118
Arabidopsis	*5 E^{-7}± 0.0135	*7 E^{-7}± 0.0132	*2 E^{-6}± 0.0133	0.0017± 0.0140		0.0639± 0.0136	*0.0003 ±0.0136	*5 E^{-6}± 0.0132	*1 E^{-7}± 0.0120
Oryza	*0.0004 ±0.0124	*0.0003 ±0.0124	0.0016± 0.0121	0.1622± 0.0134	0.0639± 0.0136		0.0068± 0.0091	0.0036± 0.0120	*0.0001± 0.0108
Maize	0.0849± 0.0113	0.0695± 0.0113	0.2048± 0.0108	0.6458± 0.0126	*0.0003± 0.0136	0.0068± 0.0091		0.3437± 0.0109	0.0756± 0.0097
Nymphaea	0.3997± 0.0108	0.3385± 0.0107	0.7493± 0.0107	0.1867± 0.0122	*5 E^{-6}± 0.0132	0.0036± 0.0120	0.3437± 0.0109		0.4664± 0.0094
Magnolia	0.8191± 0.0099	0.7317± 0.0099	0.7328± 0.0100	0.0514± 0.0118	*1 E^{-7}± 0.0120	*0.0001 ±0.0108	0.0756± 0.0097	0.4664± 0.0094	

* indicates significantly different substitution rates.

Table 4: The relative rate probability for angiosperm and gymnosperm RPB1 amino acid sequences present in this study. Psilotum was the outgroup in the rate calculations. Substitution rates are significant at $P < 0.05$.

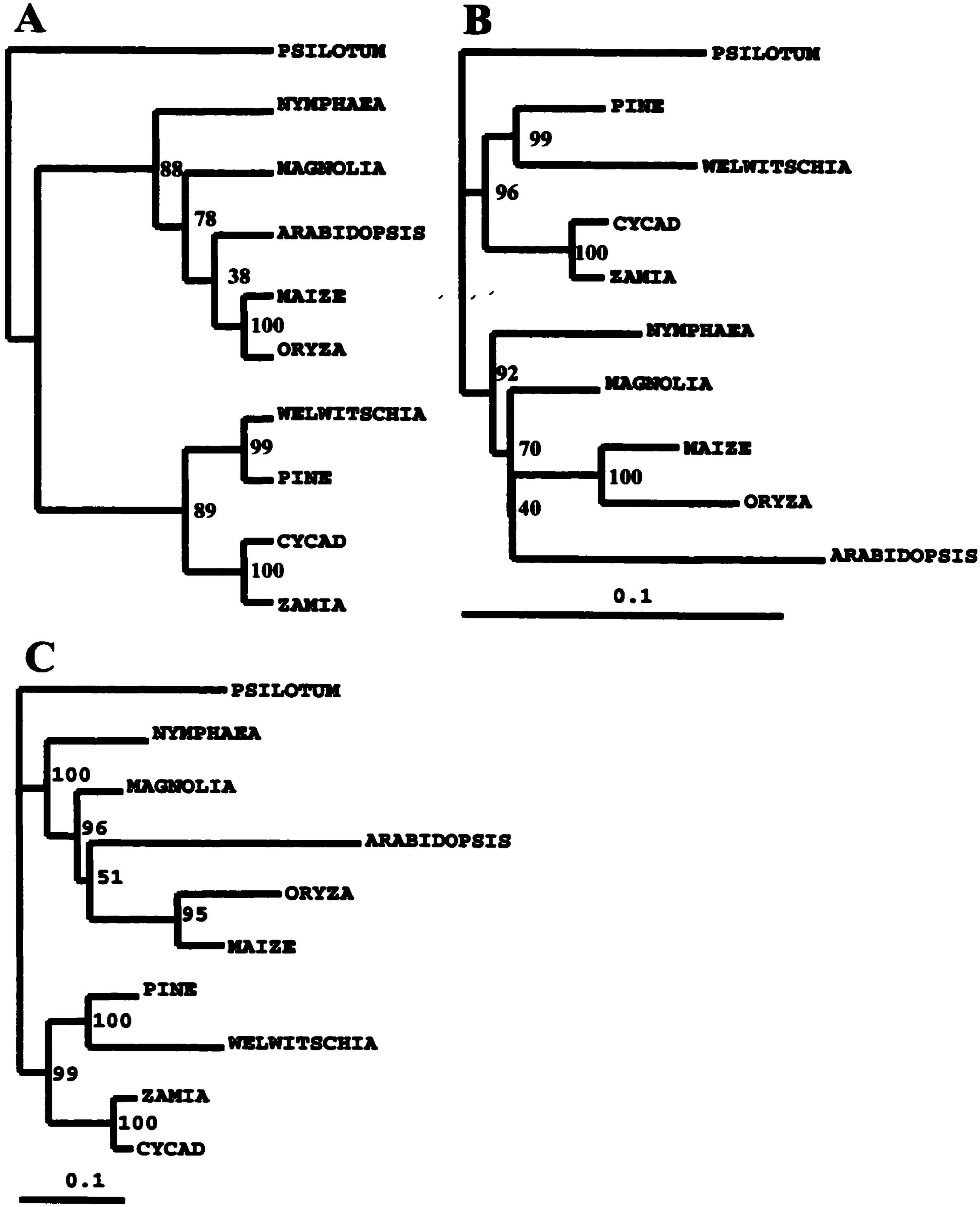
Taxon 1	Taxon 2	Probability ± standard deviation
Gymnosperms	Angiosperms	0.0042 ± 0.0075
Gymnosperms	Angiosperms without *Arabidopsis*	0.1411 ± 0.0076

Phylogenetic analyses:

All phylogenetic analyses were based on the amino acid alignment of the ten RPB1 sequences. The first three analyses utilized the complete alignment from domain A to domain G (1014 amino acids), while the last set of analyses divided the alignment into parts based on the highly conserved domains. First, we analyzed domains A to D (412 amino acids), second domains D to G (602 amino acids), third domains A to E (530 amino acids) and last, domains A to F (791 amino acids).

Psilotum was the chosen outgroup in every analysis as it branched prior to the seed plants in a maximum likelihood analysis that also contained the *Spirogyra sp.* sequence available on GenBank (accession number: U90210) as the outgroup (tree not shown). All phylogenetic analyses that utilized the complete alignment of 1014 amino acids share identical tree topologies and reveal a distinct angiosperm clade and gymnosperm clade (Fig. 6). The most basal angiosperm is *Nymphaea* with strong support values in all analyses, followed by *Magnolia* with 96% support in the maximum likelihood analysis, but only moderate bootstrap support of 70% and 78% in the neighbor-joining and parsimony analyses, respectively. *Arabidopsis* is the sister group to the monocots with weak support values in all analyses. The monocots maize and *Oryza*, group together in every analysis with very high support values between 95% and 100%. The gymnosperm clade is well supported with values no lower than 89%. Pine and *Welwitschia* form a clade sister to the *Cycad* and *Zamia* clade and the support values for these two groups are either 99% or 100%. The parsimony analysis contained 146

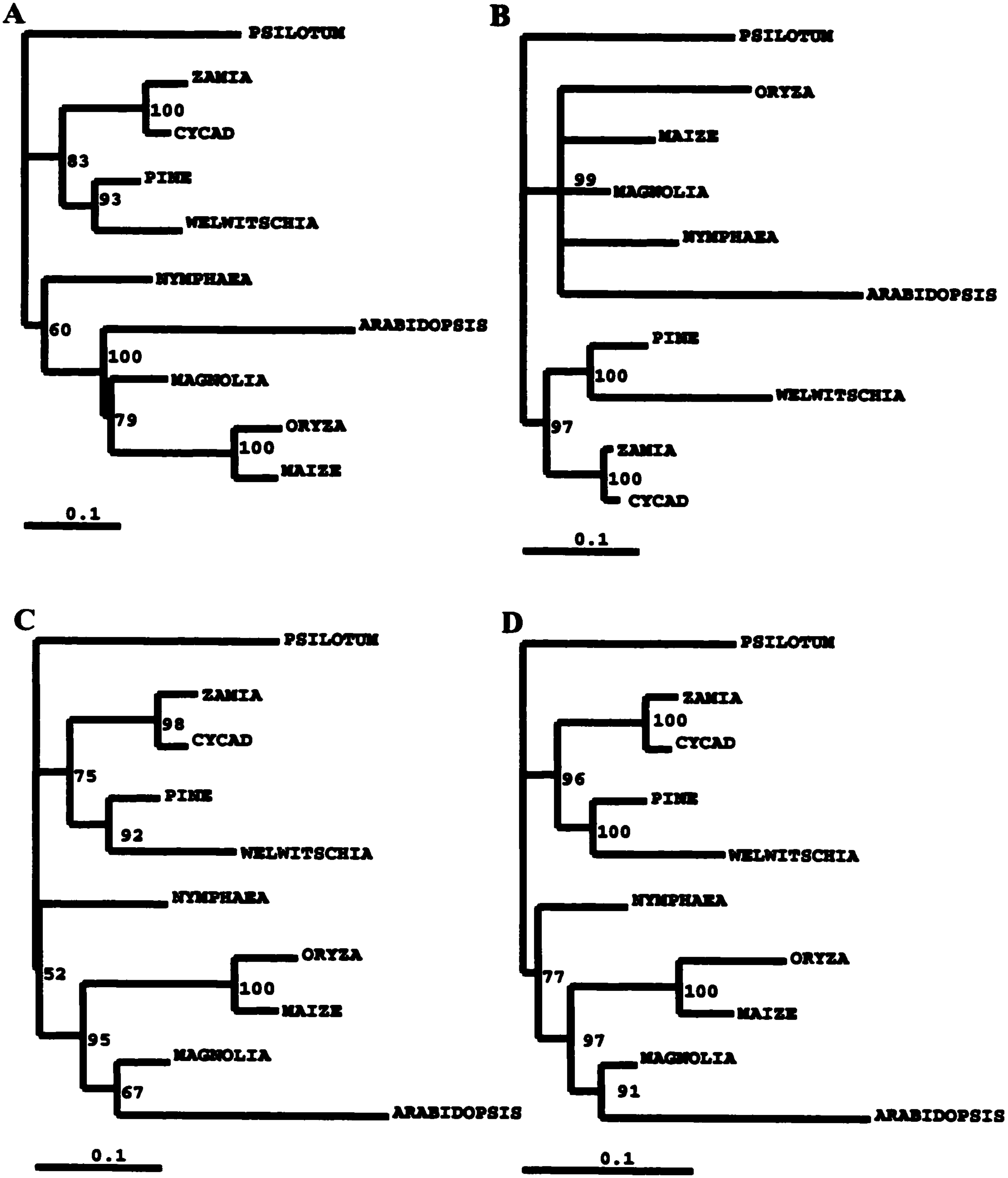
Figure 6 Phylogenetic trees reconstructed using maximum parsimony (A), neighbor-joining (B) and maximum likelihood (C) methods using all 1014 amino acids of the RPB1 alignment. A) One most parsimonious tree inferred from 146 parsimony informative sites. Support at the nodes corresponds to bootstrap values. The tree required 656 steps. Note that this tree is not to scale, i.e., the lengths of the branches are not proportional to the number of amino acid changes. B) Neighbor-joining tree where support at the nodes corresponds to bootstrap values and the scale at the base of the tree represents 0.1 substitutions per site. C) Maximum likelihood tree where values at the nodes represent puzzling support values and the scale at the base of the tree corresponds to 0.1 substitutions per site.


A
PSILOTUM
NYMPHAEA
88
MAGNOLIA
78
ARABIDOPSIS
38
MAIZE
100
ORYZA
WELWITSCHIA
99
PINE
89
CYCAD
100
ZAMIA
B
PSILOTUM
PINE
99
WELWITSCHIA
96
CYCAD
100
ZAMIA
NYMPHAEA
92
MAGNOLIA
70
MAIZE
100
ORYZA
40
ARABIDOPSIS
0.1
C
PSILOTUM
NYMPHAEA
100
MAGNOLIA
96
ARABIDOPSIS
51
ORYZA
95
MAIZE
PINE
100
WELWITSCHIA
99
ZAMIA
100
CYCAD
0.1

parsimony informative sites and needed a total of 656 steps (Fig. 6A). The neighbor-joining and maximum likelihood analyses demonstrate a very long branch for *Arabidopsis* (Fig. 6B-C). This is consistent with the calculated faster rate of evolution of its RPB1 sequence (Table 3). The maximum likelihood analysis generally showed the highest support values at all nodes with only the branch leading to *Arabidopsis* at less than 95% support (Fig. 6C). The low support value of 51% for the maximum likelihood *Arabidopsis* branch is still the highest among the three different phylogenetic analyses performed.

The maximum likelihood analyses using the amino acid alignments divided into parts (i.e., domains A to D, domains D to G, domains A to E and domains A to F) reveal some interesting information about the evolution of the RPB1 gene sequences used. Again the chosen outgroup was *Psilotum* and the analyses show a distinct angiosperm clade and gymnosperm clade (Fig. 7). The gymnosperm clade is congruent with all three analyses using the complete alignment and it is well supported. The pine and *Welwitschia* clade along with the *Zamia* and *Cycad* clade are highly supported with values of 92% and above. The analysis of domains A to D shows similar results for the topology of the angiosperms, when compared to the other three analyses utilizing the complete alignment (Figs. 6-7A). However, the A to D maximum likelihood analysis differs in its placement of *Arabidopsis* between *Nymphaea* and *Magnolia* with 100% support (Fig 7A). The basal position of *Nymphaea* and the monophyly of the monocots are consistent with the other analyses and both receive 100% support. Furthermore, the monophyly of the angiosperms receives only 60% support. The maximum likelihood analysis of domains D to G is

Figure 7 Maximum likelihood analyses using the RPB1 amino acid alignment divided into parts. Values at the nodes represent puzzling support values and the scale at the base of the trees represent 0.1 substitutions per site. A) Maximum likelihood tree inferred from the amino acid alignment of domains A to D of RPB1. B) Maximum likelihood tree inferred from the amino acid alignment of domains D to G of RPB1. C) Maximum likelihood tree inferred from the amino acid alignment of domains A to E of RPB1. D) Maximum likelihood tree inferred from the amino acid alignment of domains A to F of RPB1.


A
PSILOTUM
ZAMIA
100
CYCAD
83
PINE
93
WELWITSCHIA
NYMPHAEA
60
ARABIDOPSIS
100
MAGNOLIA
79
ORYZA
100
MAIZE
0.1
B
PSILOTUM
ORYZA
MAIZE
99
MAGNOLIA
NYMPHAEA
ARABIDOPSIS
PINE
100
WELWITSCHIA
97
ZAMIA
100
CYCAD
0.1
C
PSILOTUM
ZAMIA
98
CYCAD
75
PINE
92
WELWITSCHIA
NYMPHAEA
52
ORYZA
100
MAIZE
95
MAGNOLIA
67
ARABIDOPSIS
0.1
D
PSILOTUM
ZAMIA
100
CYCAD
96
PINE
100
WELWITSCHIA
NYMPHAEA
77
ORYZA
100
MAIZE
97
MAGNOLIA
91
ARABIDOPSIS
0.1

unable to resolve any of the relationships among the angiosperms, but the angiosperm clade is highly supported with a value of 99% (Fig. 7B). Maximum likelihood analyses of domains A to E and domains A to F share the same topology with higher support values for the tree based on the alignment comprised of domains A to F (Fig.7C-D). The angiosperm relationships are resolved, but incongruent with the results from the previous analyses because *Magnolia* and *Arabidopsis* form a clade with 67% and 91% support sister to the monocots. However, the tree topologies based on domains A to E and domains A to F are consistent with the previous RPB1 analyses presented here in their placement of *Nymphaea* at the base of the angiosperm clade (Figs. 6-7A, C-D).

The Kishino-Hasegawa test showed that all alternative hypotheses were rejected, as being significantly worse than the best tree, except for the ranalean theory hypothesis (Table 5). With a difference in the log likelihood scores of 21.60 ± 11.33, the ranalean theory was not significantly worse, at the 5% level, than the best tree. However this tree was less likely with a log likelihood score of –6187.68 compared to the best tree log likelihood score of –6166.08.

Southern hybridizations:

Each lane of the *Cycad* Southern blot shows one intense band between 1995 base pairs and 7943 base pairs (Fig. 8). The lane digested by the Eco RI restriction enzyme also shows a faint band at 6607 base pairs, due to an internal cut site in the F to G region of the gene. In maize, the DNA digested with Eco RI reveals four distinct bands between

Table 5. Alternative tree hypotheses tested using the Kishino – Hasegawa test.

Hypotheses	Log likelihood score	Difference	Standard error	Significantly worse
best tree	-6166.08	N.A.	N.A.	N.A.
ranalean theory	-6187.68	21.60	11.33	No
Gnetales as sister to Cycads	-6189.70	23.62	9.64	Yes
Gnetales placed within the Cycads	-6245.10	79.02	16.65	Yes
Gnetales as sister to all gymnosperms	-6189.25	23.17	9.88	Yes
Anthophyte Theory	-6201.61	35.53	12.06	Yes

N.A. = not applicable.

Note: Difference = log L_{tree1} - log L_{tree2}; where tree1 is the best tree and tree2 refers to any alternative hypothesis.

2818 base pairs and 5888 base pairs, while the DNA digested with either the Xba I or the Bgl II restriction enzymes show one broad band at 4467 and 8128 base pairs, respectively (Fig. 8). Due to the thickness of these bands, it is difficult to discern whether the band represents a doublet (i.e., two bands very close in size) or a single hybridization event. Another Southern hybridization, where the DNA was digested with different restriction enzymes, may clear this up. The *Welwitschia* Southern clearly demonstrates one hybridization event in each lane varying in size from 4074 base pairs to 9550 base pairs (Fig. 8). In lane 1, *Psilotum* reveals a smaller fragment at 4266 base pairs and a larger fragment at 10 471 base pairs, but in lanes 2 and 3, only one band is present at 3236 and 13 490 base pairs, respectively (Fig. 8). The *Nymphaea* hybridization events range in size from 6310 to 12 589 base pairs, while the *Magnolia* hybridization events range in size from 1259 to 12 589 base pairs (Fig. 8). Unfortunately, these two Southern blots were over-exposed and this makes it difficult to clearly see what is happening, particularly at the higher molecular weights where multiple hybridization events have occurred. The *Zamia* Southern hybridization (Fig. 8) shows 2 distinct bands in the lane digested with the Bgl II restriction enzyme ranging in size from 3311 base pairs to 5686 base pairs. The other two lanes of the *Zamia* Southern did not hybridize well, so no distinct bands are apparent. Southern hybridizations for pine were attempted a number of times with no success.

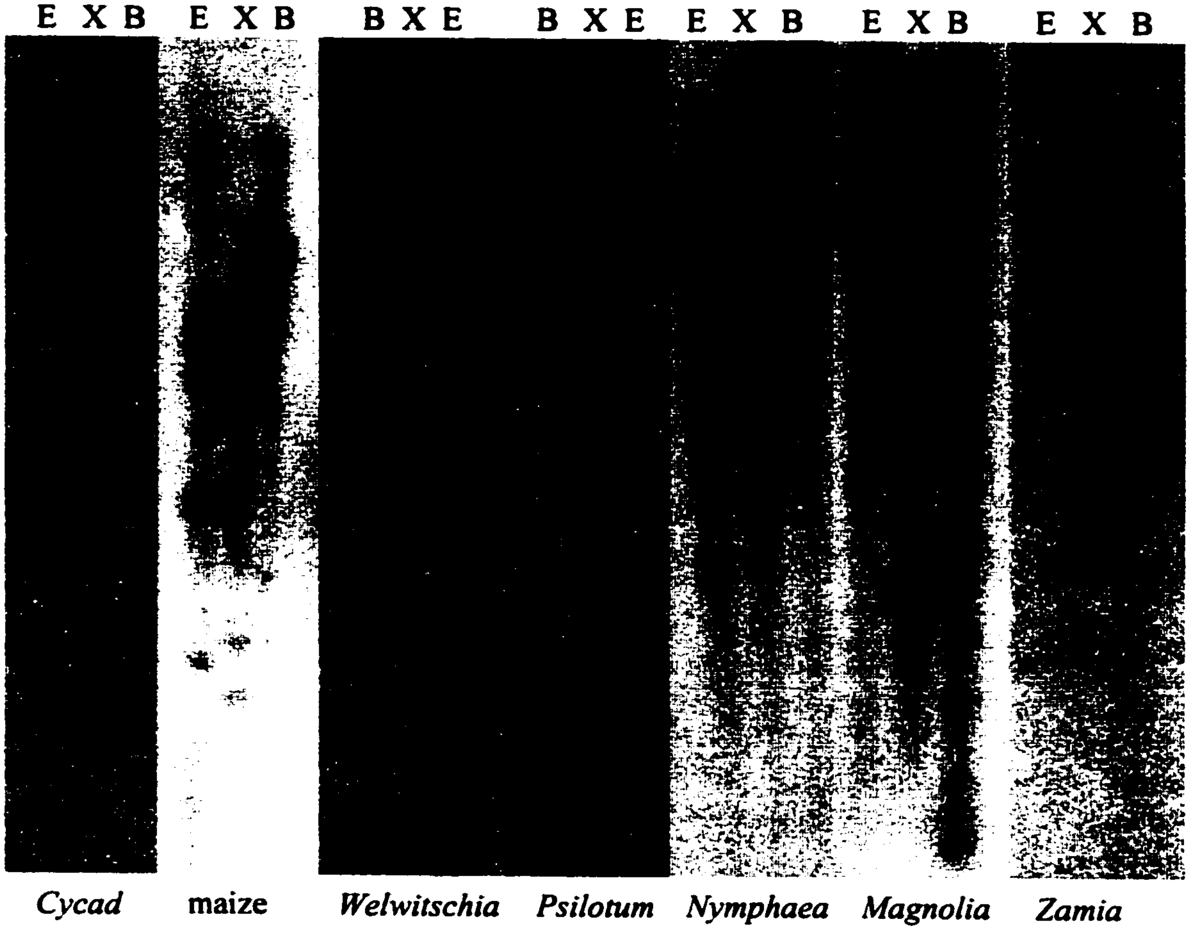


Figure 8 Southern Blot hybridizations for seven plant species. Plant genomic DNA digested with EcoR I (E), Xba I (X), and Bgl II (B) restriction enzymes. Regions F to G of RPB1 for each plant species were used as homologous probes.

Discussion

RNA Isolation from gymnosperm tissues:

The Bahloul and Burkard (1993) RNA extraction method was used to successfully isolate RNA from pine needles. The key to this method was that it contained no chaotropic salts in the isolation buffer. Commonly used chaotropic salts include guanidinium isothiocyanate and urea. Due to unidentified secondary metabolites, chaotropic salts present in the extraction buffer interfere with the resuspension of RNA (Bugos *et al.*, 1995). Polyvinylpyrrolidone (PVP) was included in the isolation buffer because it binds polyphenolics. Polyphenolic compounds, when oxidized, are known to bind irreversibly to nucleic acids and also to co-precipitate with RNA molecules; thus, inhibiting downward enzymatic activity (Chang *et al.*, 1993; Koonjul *et al.*, 1999). The authors also added the amino acid cysteine to the extraction buffer, but gave no indication as to why. Cysteine can act as a reducing agent; therefore, it might have been added to the isolation medium to prevent oxidation of the polyphenolic compounds. Gymnosperm tissues are known for their high polyphenolic and polysaccharide content. To ensure low levels of these compounds, it is necessary to collect young gymnosperm tissues during the spring and early summer months (Bahloul & Burkard, 1993; Chang *et al.*, 1993; Lonneborg & Jensen, 2000; Wang *et al.*, 2000).

Seed plant evolution:

There are a number of questions in seed plant evolution that have been asked by plant taxonomists and systematists over the past few decades such as: Are the angiosperms a monophyletic group? What is/are the closest gymnosperm(s) to the angiosperms? What are the major angiosperm lineages? What are the basal angiosperm lineages? What did the earliest angiosperms look like? When did the angiosperms arise? Why are the angiosperms so diverse?

I can't discuss every plant study done over the past few decades, so I have chosen three molecular studies that have contributed greatly to answering most of the questions listed above. They were selected based on their broad analysis of seed plants and because they were molecular studies like my own. First, there was the study by Chase *et al.* (1993) that used the chloroplast *rbc*L gene, which encodes the largest subunit of RuBisCO, to study over 500 seed plants. The plants chosen represented all of the major taxonomic groups and two separate parsimony analyses were performed from an alignment of 1428 nucleotides for 475 and 499 plant species. The major findings included *Ceratophyllum* as the first branching angiosperm, Gnetales as a highly divergent group among the seed plants, flowering plants (excluding *Ceratophyllum*) separated into two groups based on uniaperturate and triaperturate pollen types, and the analyses showed support for the monophyly of angiosperms, monocots and eudicots. Second, the study by Soltis *et al.* (1997) used the complete sequence of 18S rDNA of 223 angiosperm species. Parsimony analyses of 1850 aligned nucleotides were performed on four data sets representative of all major angiosperm groups. The results demonstrated *Amborella*, Austrobaileyaceae, Illiciaceae and Schisandraceae as the first branching angiosperms followed by Nymphaeaceae. Monocots appeared as an early monophyletic lineage, *Ceratophyllum* was found as the sister to monocots, and angiosperms were separated into two groups with uniaperturate pollen at the base of angiosperms and a large eudicot clade composed mainly of the triaperturate pollen type. Third, Chaw *et al.* (1997) used 18S rRNA sequences from a total of 65 fern, gymnosperm and angiosperm species. 1423 nucleotide sites, where 249 positions were parsimony informative, were used in neighbor-joining and parsimony analyses. The major results showed the Nymphaeaceae as the most basal angiosperm, *Magnolia* and *Ceratophyllum* in a relatively derived position nested within the angiosperms, monocots as monophyletic, Gnetales as sister to the conifers with *Ginkgo* and *Cycad* as outgroups to them, and seed plants divided into two separate groups containing angiosperms and gymnosperms, respectively.

What is most interesting about these studies is how they each support a different hypothesis for the basal-most angiosperm. The Chase *et al.* (1993) study reveals *Ceratophyllum*, a paleoherb, as the most basal angiosperm, Soltis *et al.* (1997) support the ranalean theory placing woody magnoliids at the base of the angiosperm tree, while Chaw *et al.* (1997) favor the Nymphaeaceae or water lilies, as the most primitive extant angiosperm. Conversely, all three studies agree on the monophyly of monocots, Chase *et al.* (1993) and Chaw *et al.* (1997) also agree on the monophyly of angiosperms (Note: The Soltis *et al.*, (1997) study only contained angiosperms). The analysis by Chaw *et al.* (1997) was criticized at the time for not having enough taxa, in order to explain its incongruence with other studies. Actually, this study concurs nicely with more recent studies (see pp.78-81), since it refutes the Anthophyte theory by placing Gnetales as sister to the conifers and it also demonstrates that all modern seed plants share a common ancestor; therefore, the seed evolved only once.

Each of these genes has their merits and pitfalls as phylogenetic marker molecules in plant evolutionary studies. Most notably, 18S rDNA is difficult to align because there is no clear frame of reference, since it doesn't code for protein, and the secondary structure of the RNA leads to sequencing problems and errors, but the gene was thought to be good for elucidating deep phylogenetic relationships because of the homogeneity of ribosomal RNA cistrons (Nickrendt & Soltis, 1995; Soltis *et al.*, 1997). Unfortunately, due to the rapid radiation of angiosperms, a small amount of homoplasy or error will distort some relationships because of their very short branch lengths (Soltis *et al.*, 1997). Furthermore, Soltis *et al.* (1999b) looked at 93 species of land plants and 7 green algal relatives in a parsimony analysis of 448 parsimony informative sites from an alignment containing 1739 nucleotides. They discovered that 18S rDNA had limited uses for reconstructing deep evolutionary relationships, due to its rate and pattern of evolution across land plants, even with increased character and taxon sampling (Soltis *et al.*, 1999b). *rbc*L is 3X more variable than rDNA and it contains 1.4X more phylogenetically informative sites, but it is not always maternally inherited and it is better at elucidating familial and generic level relationships than deep phylogenies (Nickrendt & Soltis, 1995; Soltis *et al.*, 1997). Thus, a new and more reliable gene needed to be used for determining the deep relationships among land plants and hopefully this gene could also answer the many questions surrounding seed plant evolution.

RPB1 as a phylogenetic marker molecule:

The gene encoding RPB1, the largest subunit of RNA polymerase II, has shown great promise at elucidating deep phylogenetic relationships. An initial study by Sidow & Thomas (1994) concluded that RPB1 was an appropriate molecule for determining early eukaryotic evolution. RPB1 has more phylogenetically informative sites than 18S rRNA, it is a protein coding gene thus it holds a high content of functional information, it contains highly conserved domains across all kingdoms of life, which simplifies PCR amplification based studies, and it is a long sequence with an appropriate evolutionary rate to resolve ancient divergences (Sidow & Thomas, 1994; Stiller & Hall, 1997; Liu et al., 1999). In 1997 & 1998, Stiller & Hall strongly rejected a relationship between the red algae and green plants using RPB1 sequences. This result was criticized by Moreira *et al.* (2000), who argued that the RPB1 sequences were susceptible to long-branch attraction in the analyses. Recently, Stiller *et al.* (2001) responded to this criticism by using RPB1 sequences from domain A to domain G and again indicated an independent origin of the red algae and green plants. Moreover, the authors examined the unique substitutions present in red algae and green plants and they were able to conclude that there was no indication that their result was due to a long-branch attraction artifact. In fact, the RPB1 sequences of red algae and green plants evolve at a slower rate than most other major eukaryotic taxa (Stiller *et al.*, 2001). Other studies using RPB1 sequences revealed that microsporidia, highly specialized eukaryotic unicells that live as parasites of other eukaryotes, are related to fungi (Hirt *et al.*, 1999) and that hexapods are more closely related to crustaceans than to myriapods (Shultz & Regier, 2000). This latter study is congruent with other molecular studies using different DNA sequences, but it conflicts with the long-held beliefs based on morphology; much like the tale involving seed plant evolution.

To attempt to answer the many questions surrounding seed plant evolution we have decided to use the RPB1 gene as a phylogenetic marker molecule. As far as we know, this is the first study to use a single nuclear protein-coding gene to elucidate seed plant relationships.

Sequence alignments:

It is quite obvious when looking at the nucleotide and amino acid RPB1 alignments for the ten plant species represented in this study that the gene is highly conserved and very easy to align (Figs. 2, 3). This high degree of conservation simplifies the designing of primers to any of the highly conserved domains present in RPB1 and it also makes this PCR based project easy to reproduce.

The first 190 amino acids contain the largest observable sequence divergence and all five indels are located in this region. All gymnosperms and the outgroup *Psilotum* retain the same indel pattern as *Nymphaea* and they all contain 3039 base pairs from domains A to G (Table 1). This pattern of indels support the basal position of the *Nymphaea* RPB1 sequence amongst the angiosperms studied.

Sequence analyses:

The G + C content of the RPB1 gene in the ten plant species studied is constant at 42% - 43% when compared to the frequency of G + C at each codon position (Fig. 4). Moreover, the total G + C content at the third codon position varies from 39% - 42% in all species when compared to the percent of A at the third codon position (Fig. 5). Therefore, the phylogenetic analyses are not biased to spurious relationships based on the G + C content of RPB1. In other words, no plant species are grouping together in the phylogenetic trees based on similar base composition.

There are multiple synonymous substitutions per nucleotide site calculated in most pairwise comparisons of the ten plant taxa (Table 2). The majority of synonymous substitutions involve the third codon position and when distantly related taxa are compared, they tend to result in saturation of the phylogenetic signal. Since there are a multiple of synonymous substitutions per nucleotide site and there is no base compositional bias, very little phylogenetic signal remains at the third codon position. When analyzing the nonsynonymous substitutions (Ka) per nucleotide site (Table 2), all values fall below 0.1092 ± 0.0073. If this Ka value is placed in the one parameter model equation $p = \frac{3}{4}(1 - e^{-1.33k})$ and we solve for p, then p = 0.1016 substitutions per site. If we subtract 0.1092 from 0.1016, we get approximately 0.01 multiple substitutions per site. Therefore, only 1% of these substitutions out of the estimated 11% of nonsynonymous substitutions per nonsynonymous site are the result of a multiple substitution and 10% of the nonsynonymous differences between the plant taxa are due to unique substitutions. Consequently, the RPB1 sequence data does not suffer from homoplasy as a result of multiple substitutions and is therefore reliable for tree reconstruction.

The relative evolutionary rates were calculated with *Psilotum* as the chosen outgroup and it was determined that *Arabidopsis* and *Oryza* evolve at significantly different rates than the other sequences studied (Table 3). Interestingly, when the data are not corrected for multiple comparisons, the *Welwitschia* RPB1 sequence also evolves at a significantly different rate as seen in many other studies (Price 1996; Samigullin *et al.*, 1999; Bowe *et al.*, 2000; Chaw *et al.*, 2000). These other studies, however, deal with multiple data sets consisting mainly of rRNA genes and plastid genome sequences that were shown to be evolving at much faster rates and in some cases, sensitive to rooting of the phylogenetic trees (Bowe *et al.*, 2000; Chaw *et al.*, 2000). The number of nucleotide substitutions observed among the Gnetales is very large and signifies a long period of divergence similar in magnitude to the distances between monocots and eudicots (Price, 1996). Vastly different evolutionary rates may result in a phylogenetic tree reconstruction artifact known as long-branch attraction. Since the *Arabidopsis* and *Oryza* RPB1 sequences evolve at significantly different rates than the other RPB1 sequences, they may pose a problem in the phylogenetic trees if they group with sequences based on similar rates of evolution. When the sequences are grouped into two lineages, angiosperms and gymnosperms, the rate of evolution of these two groups is significantly different (Table 4). This is a direct result of the more divergent *Arabidopsis* RPB1 sequence because when *Arabidopsis* is excluded from the angiosperm lineage, the rate of evolution of these two groups is no longer statistically significant (Table 4). RPB1 has a conservative rate of evolution in most plant species presented here and thus evolves at an appropriate rate for inferring deep evolutionary relationships.

Phylogenetic analyses:

All three phylogenetic tree reconstruction methods (i.e., maximum parsimony, neighbor-joining, and maximum likelihood) based on the RPB1 amino acid alignment from domain A to domain G are congruent with similar support values at all nodes (Fig. 6). The trees show two distinct clades containing the angiosperm and gymnosperm species. The gymnosperm clade is composed of a pine and *Welwitschia* group sister to a *Cycad* and *Zamia* group. The angiosperm clade places *Nymphaea* as the most primitive extant angiosperm followed by *Magnolia*. *Arabidopsis* is the only branch with consistently low support values and it is located between *Magnolia* and the monocot clade containing maize and *Oryza*. The low support values, in all three analyses for the branch leading to *Arabidopsis*, is due to its higher rate of evolution resulting in a much greater RPB1 sequence divergence.

The RPB1 analyses reported here do not support the Anthophyte Theory which places the Gnetales, a group comprised of three extant genera known as *Gnetum*, *Welwitschia* and *Ephedra*, as the closest extant relative of the angiosperms. Most support for the Anthophyte Theory is based on morphological studies and the term "anthophyte" was coined to describe the shared possession of flower-like reproductive structures between the angiosperms and the Gnetales. A number of recent molecular studies have put an end to the Anthophyte Theory. First, Samigullin and colleagues (1999) used the chloroplast gene *rpo*C1, which encodes the b'-subunit of chloroplast DNA dependent RNA polymerase. They showed that the angiosperms and gymnosperms form two monophyletic clades with strong support and that *Gnetum* and *Pinus* group together in neighbor-joining and maximum likelihood analyses. The alignment of 1654 base pairs also displayed a deletion at position 125 – 126 common to all angiosperms and an insertion at position 116 common to *Gnetum* and *Pinus* to establish further disagreement with the Anthophyte hypothesis. Second, a study by Winter *et al.* (1999) also revealed strong support for a *Gnetum* and pine clade when MADS-box genes were utilized. Thirteen different homologs of MADS-type floral homeotic genes were studied from *Gnetum gnemon*, where five of these genes fall into monophyletic clades containing orthologous genes from conifers and flowering plants. Within these clades, *Gnetum* always groups with the conifers as it shares a high degree of sequence similarity with the conifer sequences. Two of these five clades were the orthologous floral homeotic B and C functional genes. B and C genes together specify stamen formation, while C genes alone specify the formation of carpels (Winter *et al.*, 1999). Consequently, the authors concluded that the presence of orthologous floral homeotic B and C genes in gymnosperms suggests that the system specific for reproductive organ identity in angiosperms (i.e., stamen and carpel formation) was retained from a common ancestor of seed plants that had a similar system approximately 300 MYA. Third, Hansen *et al.* (1999) also show *Gnetum* branching with its *Pinus* homologs in a study containing a combined data set utilizing a final alignment of 9149 sites. Of these sites, 6 362 were invariant and 1 167 were autapomorphic. However, a weakness of this study is that no other gymnosperms were present and no basal angiosperms were studied either. A fourth study by Chaw and colleagues (2000) proposed a new theory, called the "gnepines" hypothesis, to explain the sister relationship between the Pinaceae and the Gnetales. They analyzed a data set from all three plant genomes and their results revealed paraphyly among the conifers, as the Gnetales allied with the Pinaceae, and a monophyletic gymnosperm clade with Cycads as the most basal group. Finally, Bowe *et al.* (2000) examined a total of four genes separately and then combined these genes into a multiple data set, which now contained genes from all three plant genomes. The unrooted *cox*1 and *atp*A tree topologies were identical with angiosperms, Gnetales and Cycads as monophyletic groups and the conifers were paraphyletic, due to the sister relationship between the Gnetales and the Pinaceae. The unrooted analysis of the *rbc*L gene agreed with the *cox*1 and *atp*A results, but the 18S rDNA analysis placed Gnetales sister to a monophyletic conifer group. Rooted trees of the combined data set support the monophyly of angiosperms, gymnosperms, Cycads and Gnetales, Cycads were positioned at the base of the tree and Gnetales and Pinaceae formed a clade with moderate support. In all cases, the anthophyte theory was never supported, but whether or not the Gnetales are sister to the conifers or derived from within the conifers and sister to the Pinaceae is still unresolved. Donoghue and Doyle (2000) reviewed all of the above-mentioned studies and they are skeptical of the Gnetales being derived from within the conifers.

They concluded that more molecular studies need to be conducted to explore the conflicts surrounding the Gnetales and that more attention has to be paid to morphological characters in order to enhance the credibility of future studies. It is now clear that no living non-flowering seed plants are closely related to the angiosperms, so determining their origins is hopefully hidden somewhere in the fossil record.

The analyses of the ten plant RPB1 sequences presented here concur with all recent findings in the basal position of *Nymphaea* among flowering plants. The long-standing debate over the most basal angiosperm lineages is not over, but much has been learned over the passed few years. In 1999, four independent studies all concluded that the ANITA clades, comprised of *Amborella* (A), *Nymphaea* (N), and an Illiciaceae, Schisandraceae, Trimeniaceae and Austrobaileyaceae (ITA) group, are the most primitive of extant angiosperms (Matthews & Donoghue, 1999; Parkinson *et al.*, 1999; Qiu *et al.*, 1999; Soltis *et al.*, 1999a). In all studies, *Amborella* was the basal-most branch followed by *Nymphaea*. Matthews & Donoghue (1999) used duplicate phytochrome genes (*PHY*A and *PHY*C) in parsimony analyses to root the angiosperms to avoid any bias contributed from highly divergent outgroups, while Soltis and colleagues (1999a) analyzed 560 angiosperms and 7 non-flowering seed plants in parsimony analyses of *rbc*L, *atp*B and 18S rRNA genes. They generated a well supported and well resolved tree congruent with other studies for overall angiosperm phylogeny. Qiu *et al.* (1999) used a total of five DNA sequences with representatives from all three plant genomes for 105 plant species. They also observed an indel found in the *mat*R gene that further supports the basal position of the ANITA clades, as an 18 base pair deletion is found in all angiosperms, except the ANITA clades, and this deletion is not present in the gymnosperms studied either. The extensive sampling of taxa and characters are the real strength of this study as it uses multiple genes of different function from all three genomes to reduce homoplasy generated by gene, function or genome specific phenomena; such as, rate heterogeneity, GC content bias, RNA editing and protein structural constraints. Parkinson and colleagues (1999) also studied five genes, representing all three plant genomes, in parsimony and maximum likelihood analyses. Their analyses of 45 angiosperms and 6 gymnosperms also tested alternative topologies for the earliest branching angiosperms using the maximum likelihood Kishino-Hasegawa test. They determined that all alternative hypotheses are rejected except for an *Amborella* and/or Nymphaeales root of flowering plants. Likewise, Barkman *et al.* (2000) resolved *Nymphaea* and *Amborella* together as the first branching angiosperms in a study of six and nine genes for different subsets of taxa. Lastly, Graham and Olmstead (2000) utilized 17 chloroplast genes in parsimony analyses and agreed with the monophyly of angiosperms and that *Nymphaea* and the ITA group were near the base of the angiosperm tree. However, when they added *Amborella* to their core analysis, *Amborella* was strongly rejected as the root in favor of *Cabomba*, a water lily, but when *Amborella* and *Nymphaea odorata* were added together, the tree rooted with *Amborella* while *Nymphaea* and *Cabomba* formed a group sister to *Amborella*. Consequently, they advise that it is premature to place any confidence in the rooting of the angiosperm tree with *Amborella*. All of these studies, except for Barkman *et al.* (2000), are reviewed in Kuzoff and Gasser (2000).

What can we conclude from all of these analyses about the origins of angiosperms? Most importantly, earlier hypotheses about the origins of extant angiosperms like the ranalean theory, Chloranthaceae and Casuarinaceae hypotheses should no longer be regarded as candidates for the earliest angiosperms. In addition, the phylogenies point to a woody growth habit for the last common ancestor of all angiosperms, since the ITA clade and *Amborella* are all woody plants suggesting that the herbaceous nature of the Nymphaeales is derived. Furthermore, the history of the Nymphaeales in the fossil record now extends to the early Cretaceous (115-125 MYA) with the report of explicit evidence of fossil floral structures and pollen (Friis *et al.*, 2001). This places the Nymphaeales into the oldest fossil assemblages that contain only one other angiosperm family, the Chloranthaceae (Friis *et al.*, 2001). These molecular and paleobotanical findings change how we look at the evolution of the angiosperms and they provide valuable insights into determining primitive characters for morphology-based studies, which are needed to finally resolve this "abominable mystery".

To further investigate the observed higher sequence divergence in the first 190 amino acids of the RPB1 sequences, the alignment was divided into several parts containing domains A to D, domains D to G, domains A to E and domains A to F. Maximum likelihood analyses were performed on these data sets (Fig. 7) and the RPB1 genes have revealed some interesting information about their evolution in land plants. First of all, the gymnosperm lineages are resolved in all analyses, so enough phylogenetically informative signal is present to resolve the relationships among these distantly related seed plants. Secondly, the angiosperm relationships are completely unresolved in the tree based on the D to G regions of RPB1, but resolved in the analyses based on domains A to D, domains A to E and domains A to F. Consequently, the higher observed sequence divergence in the first 190 amino acids, which includes five small indels, holds more information for elucidating relationships among flowering plants than the D to G regions of RPB1. The A to D maximum likelihood tree does misrepresent the *Arabidopsis* branch when compared to the maximum parsimony, neighbor-joining, and maximum likelihood trees containing all 1014 amino acids (Figs. 6-7A). This long-branch attraction artifact is directly related to the higher evolutionary rate of *Arabidopsis*. The A to E and A to F maximum likelihood trees also show inconsistencies when compared to the analyses based on the alignment from domains A to G (Figs. 6, 7C-D). *Arabidopsis* again is problematic in these latter analyses as it branches together with *Magnolia* forming a clade sister to the monocots. These findings indicate that the complete alignment of RPB1 containing 1014 amino acids is needed to infer the correct phylogeny among seed plants in deep evolutionary studies. Similarly, Graham & Olmstead (2000) concluded in their study of 17 chloroplast genes that slowly evolving regions require more total characters to resolve deep evolutionary relationships separated by short internal branches. Thus, RPB1's conservative rate of evolution make it a good phylogenetic marker for elucidating deep relationships among land plants, but more total characters are needed to ensure a sufficient resolution.

Alternative hypotheses for plant evolution were tested using the Kishino-Hasegawa test. Surprisingly, the ranalean theory, which places the order Magnoliales or the Winteraceae family as the most primitive angiosperms, was not rejected as significantly worse than the best tree (Table 5). However, the ranalean hypothesis is less likely and with the addition of *Amborella* to the data set, the theory would probably be rejected because more weight would be placed at the base of the angiosperm tree. The other hypotheses were rejected as significantly worse than the best tree (Table 5). These results refute the Anthophyte theory and they do not support any other placement of the Gnetales in the gymnosperm clade, except for when it is grouped with pine. Therefore, further tests that include *Ginkgo* and other conifers would complete the investigation into the gnepines hypothesis proposed by Chaw *et al.* (2000).

Southern hybridizations:

The *Arabidopsis* and *Oryza* genome sequencing projects have so far revealed a single RPB1 gene in each of these species; therefore, this gene is most likely single copy in both of these species. According to the results of the Southern hybridizations (Fig. 8), *Psilotum*, *Welwitschia* and *Cycad* all possess one copy of the RPB1 gene. Consequently, homologous genes are being compared during the phylogenetic analyses. In the case of *Zamia*, *Nymphaea*, *Magnolia* and maize, there is the possibility that different copies of the RPB1 gene are being compared as more than a single copy of the RPB1 gene is present in these plants. However, when compiling the consensus sequences for each of these plants, allelic variation was the only variation ever noticed. Since the consensus sequences for *Zamia*, *Nymphaea*, *Magnolia* and maize were based on 3 – 5 clones and each clone was PCR amplified in the same manner, orthologous RPB1 genes were most likely compared during the phylogenetic analyses.

Future directions:

This study is a good initial investigation into the usefulness of RPB1 for inferring land plant phylogeny. The basal position of *Nymphaea* followed by *Magnolia* was confirmed along with the monophyly of angiosperms and gymnosperms. Unfortunately, I was unsuccessful in the isolation of RNA from *Ginkgo* or another conifer, like *Araucaria*. As a result, I was unable to address the "gnepines" hypothesis proposed by Chaw (2000) because in my analyses *Welwitschia* and pine form a clade, but no other conifers were present. I could not confirm the basal position of Cycads among the gymnosperms either because *Ginkgo* was missing from the analyses. Furthermore, another dicot may be a better exemplar species than the fast evolving *Arabidopsis* sequence available in GenBank. I would also have liked the RPB1 sequence for *Amborella* in the analyses to see if *Amborella* really is the basal-most extant angiosperm or if a clade containing both *Amborella* and *Nymphaea* is sister to all other angiosperms. Lastly, a genomic Southern hybridization for pine would have completed the analysis of the RPB1 copy number in all plant species studied.

This study may be criticized because so few taxa are represented and because the inference for land plant phylogeny is based on a single gene tree. However, as far as the Gnetales are concerned, one representative is enough because both morphological and molecular studies have strongly indicated that the Gnetales are a monophyletic group (Hansen et al., 1999). The strength of this study is the number of characters available for the analyses. It has been shown that the percent of supported nodes is positively correlated with the number of characters and negatively correlated to the number of taxa utilized (Bremer et al. 1999). In regards to the inferred phylogeny being based on a single gene, it is possible for gene duplication and horizontal gene transfer to produce a gene tree different from the species tree (Slowinski & Page, 1999). Gene duplication causes conflict in a phylogenetic analysis when paralogous sequences are compared and horizontal gene transfer obscures species phylogeny because gene sequences may cross the species boundary (Slowinski & Page, 1999). It is also generally accepted that gene transfer leads to the formation of pseudogenes, which have a higher substitution rate than normally expressed genes because all mutations are selectively neutral (Laroche *et al.*, 1997). To bypass this pseudogene issue, we performed RT-PCR on total messenger RNA, so only expressed RPB1 genes were PCR amplified. Analyses of multiple data sets are now popular and researchers are apt to compare genes from all three plant genomes to try and reduce any systematic error or bias. However, concatenating gene sequences from all three plant genomes should be done with caution as nuclear genes evolve the fastest, followed by chloroplast genes and then mitochondrial genes. This rate heterogeneity may lead to a well supported tree topology based on phylogenetic artifact instead of historical signal (Stiller *et al.*, 2001).

Conclusions:

It is apparent that molecular studies do have their drawbacks. For instance, molecular data often disagree, due to phylogenetic artifacts, like long-branch attraction or incorrect rooting, that are caused by rate heterogeneity and homoplasy (Doyle, 1998).

This is confounded in the study of flowering plants because of the rapid diversification of this group during the Siluro–Devonian period. Morphology based studies also have their problems because the fossil record is incomplete and inadequate for most plants and the tremendous diversity between similar life cycle phases of many plant groups make it difficult to determine character homologies (Bateman *et al.*, 1998; Kenrick, 2000). However, molecular studies have been useful in their ability to change the way we look at plant evolution by suggesting new relationships among extant plants and this knowledge helps morphology based studies in determining primitive character states. Both molecular studies and cladistic analyses based on morphology will be needed to answer the many remaining questions about land plant evolution.

RPB1 has shown great promise in this study in its ability to resolve deep evolutionary relationships among the seed plants. RPB1 generates consistent results with maximum parsimony, neighbor-joining and maximum likelihood analyses and all of these analyses are congruent with the most recent findings of other laboratories. Our substitution analyses have shown that the nonsynonymous sites of RPB1 evolve relatively slowly and that only 10% of such sites have been subject to multiple substitutions. They also showed that the RPB1 gene of most plant species studied, with the exception of *Arabidopsis* and *Oryza*, evolve at a similar rate. In addition, we were able to address our first question on whether *Nymphaea* or *Magnolia* held a more basal position in the flowering plant phylogenetic tree, but due to missing taxa, we were unable to address the "gnepines" hypothesis proposed by Chaw (2000). Finally, our hypothesis was accepted, as RPB1 is a useful phylogenetic marker molecule for inferring land plant evolution.

References

Allison, L.A., Moyle, M., Shales, M. & Ingles, C.J. (1985) Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. *Cell.* 42:599-610.

Bahloul, M. & Burkard, G. (1993) An improved method for the isolation of total RNA from spruce tissues. *Plant Molecular Biology Reporter* 11:212-215.

Barkman, T.J., Chenery, G., McNeal, J.R., Lyons-Weiler, J., Ellisens, W.J., Moore, G., Wolfe, A.D. & dePamphilis, C.W. (2000) Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. *Proc. Natl. Acad. Sci. USA.* 97:13166-13171.

Bateman, R.M., Crane, P.R., DiMichele, W.A., Kenrick, P.R., Rowe, N.P., Speck, T. & Stein, W.E. (1998) Early evolution of land plants: phylogeny, physiology, and the ecology of the primary terrestrial radiation. *Annu. Rev. Ecol. Syst.* 29:263-292.

Bhattacharya, D. & Medlin, L. (1998) Algal phylogeny and the origin of land plants. *Plant Physiol.* 116:9-15.

Bowe, L.M., Coat, G. & dePamphilis, C.W. (2000) Phylogeny of seed plants based on all three genomic compartments: Extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. *Proc. Natl. Acad. Sci. USA.* 97:4092-4097.

Bremer, B., Jansen, R.K., Oxelman, B., Backlund, M., Lantz, H. & Kim, K-J. (1999) More characters or more taxa for a robust phylogeny-Case study from the coffee family (Rubiaceae). *Syst. Biol.* 48:413-435.

Bremer, K. (2000) Early Cretaceous lineages of monocot flowering plants. *Proc. Natl. Acad. Sci. USA.* 97:4707-4711.

Bugos, R.C., Chiang, V.L., Zhang, X-H., Campbell, E.R., Podila, G.K. & Campbell, W.H. (1995) RNA isolation from plant tissues recalcitrant to extraction in guanidine. *BioTechniques* 19:734-737.

Chang, S., Puryear, J. & Cairney, J. (1993) A simple and efficient method for isolating RNA from Pine trees. *Plant Molecular Biology Reporter* 11:113-116.

Chase, M.W., Soltis, D.E., Olmstead, R.G., Morgan, D., Les, D.H., Mishler, B.D., Duvall, M.R., Price, R.A., Hills, H.G., Qiu, Y-L., Kron, K.A., Rettig, J.H., Conti, E., Palmer, J.D., Manhart, J.R., Systma, K.J., Michaels, H.J., Kress, W.J., Karol, K.G., Clark, W.D., Hedren, M., Gaut, B.S., Jansen, R.K., Kim, K-J., Wimpee, C.F., Smith, J.F., Furnier, G.R., Strauss, S.H., Xiang, Q-Y., Plunkett, G.M., Soltis, P.S., Swensen, S.M., Williams, S.E., Gadek, P.A., Quinn, C.J., Eguiarte, L.E., Golenberg, E., Learn, G.H. Jr., Graham, S.W., Barrett, S.C.H., Dayanandan, S. & Albert, V.A. (1993) Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbc*L. *Ann. Missouri Bot. Gard.* 80:528-580.

Chaw, S-M., Zharkikh, A., Sung, H-M., Lau, T-C. & Li, W-H. (1997) Molecular phylogeny of extant gymnosperms and seed plant evolution: Analysis of nuclear 18S rRNA sequences. *Mol. Biol. Evol.* 14:56-68.

Chaw, S-M., Parkinson, C.L., Cheng, Y., Vincent, T.M. & Palmer, J.D. (2000) Seed plant phylogeny inferred from all three plant genomes: Monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc. Natl. Acad. Sci. USA.* 97:4086-4091.

Chomcyzynski & Sacchi (1987) *Analytical Biochem.* 162:156-159.

Crane, P.R., Friis, E.M. & Pedersen, K.R. (1995) The origin and early diversification of angiosperms. *Nature* 374:27-33.

Crepet, W.L. (1998) The abominable mystery. *Science.* 282:1653-1654.

Cummings, M.P., Otto, S.P. & Wakeley, J. (1999) Genes and other samples of DNA sequence data for phylogenetic inference. *Biol. Bull.* 196:345-350.

Dallimore, W. & Jackson, A.B. (1948) A handbook of coniferae including Ginkgoaceae. Edward Arnold Publishers Ltd. Great Britain. pp. 514-521.

De Franceschi, D. & Vozenin-Serra, C. (2000) Origine du *Gingko biloba* L. Approche phylogenetique. *C.R. Acad. Sci. Paris, Sciences de la vie/Life Sciences.* 323:583-592.

Denton, A.L., McConaughy, B.L. & Hall, B.D. (1998) Usefulness of RNA polymerase II coding sequences for estimation of green plant phylogeny. *Mol. Biol. Evol.* 15:1082-1085.

Donoghue, M.J. & Doyle, J.A. (2000) Seed plant phylogeny: Demise of the anthophyte hypothesis? *Current Biology* 10:R106-R109.

Doyle, J.A. (1996) Seed plant phylogeny and the relationships of Gnetales. *Int. J. Plant Sci.* 157:S3-S39.

Doyle, J.A. (1998) Phylogeny of vascular plants. *Annu. Rev. Ecol. Syst.* 29:567-599.

Felsenstein, J. (1993) PHYLIP (phylogeny inference package). Version 3.5. Distributed by the author (ftp://evolution.genetics.washington.edu/), Department of Genetics, University of Washington, Seattle.

Friedman, W.E. (1996) Introduction to biology and evolution of the Gnetales. *Int. J. Plant Sci.* 157:S1-S2.

Friis, E.M., Pedersen, K.R. & Crane, P.R. (2001) Fossil evidence of water lilies (Nymphaeales) in the Early Cretaceous. *Nature*. 410:357-360.

Frohlich, M.W. (1999) MADS about Gnetales. *Proc. Natl. Acad. Sci. USA*. 96:8811-8813.

Graham, L.E., Cook, M.E. & Busse, J.S. (2000) The origin of plants: Body plan changes contributing to a major evolutionary radiation. *Proc. Natl. Acad. Sci. USA*. 97:4535-4540.

Graham, S.W. & Olmstead, R.G. (2000) Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *American Journal of Botany* 87:1712-1730.

Graybeal, A. (1998) Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* 47:9-17.

Hamby, R.K., Sims, L., Issel, L. & Zimmer, E. (1988) Direct ribosomal RNA sequencing: Optimization of extraction and sequencing methods for work with higher plants. *Plant Molecular Biology Reporter* 6:175-192.

Hansen, A., Hansmann, S., Samigullin, T., Antonov, A. & Martin, W. (1999) *Gnetum* and the angiosperms: Molecular evidence that their shared morphological characters are convergent, rather than homologous. *Mol. Biol. Evol.* 16:1006-1009.

Hirt, R.P., Logsdon, J.M., Healy, B., Dorsy, M.W., Doolittle, W.F. & Embley, T.M. (1999) Microsporidia are related to fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. *Proc. Natl. Acad. Sci. USA*. 96:580-585.

Holm, S. (1979) A simple sequencially rejective multiple test procedure. *Scand. J. Stat.* 6:65-70.

Iwabe, N., Kuma, K-I., Kishino, H., Hasegawa, M. & Miyata, T. (1991) Evolution of RNA polymerases and branching patterns of the three major groups of archaebacteria. *J. Mol. Evol.* 32:70-78.

Kenrick, P. & Crane, P.R. (1997) The origin and early evolution of plants on land. *Nature* 389:33-39.

Kenrick, P. (2000) The relationships of vascular plants. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*. 355:847-855.

Klenk, H-P., Palm, P. & Zillig, W. (1994) DNA-dependent RNA polymerases as phylogenetic marker molecules. *System. Appl. Microbiol.* 16:638-647.

Klenk, H-P., Zillig, W., Lanzendorfer, M., Grampp, B. & Palm, P. (1995) Location of protist lineages in a phylogenetic tree inferred from sequences of DNA-dependent RNA polymerases. *Arch. Protistenkd.* 145:221-230.

Koonjul, P.K., Brandt, W.F., Farrant, J.M. & Lindsey, G.G. (1999) Inclusion of polyvinylpyrrolidone in the polymerase chain reaction reverses the inhibitory effects of polyphenolic contamination of RNA. *Nucleic Acids Research.* 27:915-916.

Kumar, S., Tamura, K., & Nei, M. (1993) MEGA: molecular evolutionary genetics analysis. Version 1.02. The Pennsylvania State University, University Park.

Kuzoff, R.K. & Gasser, C.S. (2000) Recent progress in reconstructing angiosperm phylogeny. *Trends in Plant Science* 5:330-336.

Laroche, J., Li, P., Maggia, L. & Bousquet, J. (1997) Molecular evolution of angiosperm mitochondrial introns and exons. *Proc. Natl. Acad. Sci. USA.* 94:5722-5727.

Lewis, L.A., Mishler, B.D. & Vilgalys, R. (1997) Phylogenetic relationships of the Liverworts (Hepaticae), a basal embryophyte lineage, inferred from nucleotide sequence data of the chloroplast gene *rbc*L. *Mol. Phylo. Evol.* 7:377-393.

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

Liu, Y., Whelen, S. & Hall, B.D. (1999) Phylogenetic relationships among ascomycetes: Evidence from an RNA polymerase II subunit. *Mol. Biol. Evol.* 16:1799-1808.

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

Loconte, H. (1996) Chapter 10, "Comparison of alternative hypotheses for the origin of angiosperms" in Flowering Plant Origin, Evolution and Phylogeny. Eds. Taylor, D.W. & Hickey, L.J. Chapman & Hall. N.Y., U.S.A.

Lonneborg, A. & Jensen, M. (2000) Reliable and reproducible method to extract high-quality RNA from plant tissues rich in secondary metabolites. *BioTechniques.* 29:714-718.

Matthews, S. & Donoghue, M.J. (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science.* 286:947-950.

Memet, S., Saurin, W. & Sentenac, A. (1988) RNA polymerases B and C are more closely related to each other than to RNA polymerase A. *J. Biol. Chem.* 263:10048-10051.

Miller, C.N. Jr. (1999) Implications of fossil conifers for the phylogenetic relationships of living families. *Bot. Rev.* 65:239-277.

Mishler, B.D., Lewis, L.A., Buchheim, M.A., Renzaglia, K.S., Garbary, D.J., Delwiche, C.F., Zechman, F.W., Kantz, T.S. & Chapman, R.L. (1994) Phylogenetic relationships of the "Green Algae" and "Bryophytes". *Ann. Missouri Bot. Gard.* 81:451-483.

Moreira, D., Le Guyader, H. & Philippe, H. (2000) The origin of red algae and the evolution of chloroplasts. *Nature.* 405:69-72.

Nawrath, C., Schell, J. & Koncz, C. (1990) Homologous domains of the largest subunit of eukaryotic RNA polymerase II are conserved in plants. *Mol. Gen. Genet.* 223:65-75.

Nickrent, D.L. & Soltis, D.E. (1995) A comparison of angiosperm phylogenies from nuclear 18S rDNA and *rbc*L sequences. *Ann. Missouri Bot. Gard.* 82:208-234.

Oxelman, B. & Bremer, B. (2000) Discovery of paralogous nuclear gene sequences coding for the second-largest subunit of RNA polymerase II (RPB2) and their phylogenetic utility in Gentianales of the Asterids. *Mol. Biol. Evol.* 17:1131-1145.

Page, R.D.M. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences.* 12:357-358.

Palenik, B. (1992) Polymerase evolution and organism evolution. *Current Biology* 2:931-936.

Parkinson, C.L., Adams, K.L. & Palmer, J.D. (1999) Multigene analyses identify the three earliest lineages of extant flowering plants. *Current Biology* 9:1485-1488.

Price, R.A. (1996) Systematics of the Gnetales: A review of morphological and molecular evidence. *Int. J. Plant Sci.* 157:S40-S49.

Puhler, G., Leffers, H., Gropp, F., Palm, P., Klenk, H-P., Lottspeich, F., Garrett, R.A. & Zillig, W. (1989) Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. *Proc. Natl. Acad. Sci. USA.* 86:4569-4573.

Qiu, Y-L., Lee, J., Bernasconi-Quadroni, F., Soltis, D.E., Soltis, P.S., Zanis, M., Zimmer, E.A., Chen, Z., Savolainen, V. & Chase, M.W. (1999) The earliest angiosperms: from mitochondrial, plastid, and nuclear genomes. *Nature.* 402:404-407.

Qiu, Y-L., Cho, Y. & Cox, C. (1998) The gain of three mitochondrial introns identifies liverworts as the earliest land plants. *Nature.* 394:671-674.

Qiu, Y-L. & Palmer, J.D. (1999) Phylogeny of early land plants: insights from genes and genomes. *Trends in Plant Science.* 4:26-30.

Renzaglia, K.S., Duff, R.J., Nickrendt, D.L. & Garbary, D.J. (2000) Vegetative and reproductive innovations of early land plants: implications for a unified phylogeny.

Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences. 355:769-793.

Robinson, M., Gout, M., Gauthier, C. & Mouchiroud, D. (1998) Sensitivity of the relative-rate test to taxonomic sampling. *Mol. Biol. Evol.* 15:1091-1098.

Rothwell, G.W. (1999) Fossils and ferns in the resolution of land plant phylogeny. *Bot. Rev.* 65:188-218.

Samigullin, T.Kh., Martin, W.F., Troitsky, A.V. & Antonov, A.S. (1999) Molecular data from the chloroplast *rpo*C1 gene suggest a deep and distinct dichotomy of contemporary spermatophytes into two monophyla: Gymnosperms (including Gnetales) and angiosperms. *J. Mol. Evol.* 49:310-315.

Shultz, J.W. & Regier, J.C. (2000) Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*. 267:1011-1019.

Sidow, A. & Thomas, W.K. (1994) A molecular evolutionary framework for eukaryotic model organisms. *Current Biology*. 4:596-603.

Slowinski, J.B. & Page, R.D.M. (1999) How should species phylogenies be inferred from sequence data? *Syst. Biol.* 48:814-825.

Soltis, E.D., Soltis, P.S., Nickrent, D.L., Johnson, L.A., Hahn, W.J., Hoot, S.B., Sweere, J.A., Kuzoff, R.K., Kron, K.A., Chase, M.W., Swensen, S.M., Zimmer, E.A., Chaw, S-M., Gillespie, L.J., Kress, W.J. & Sytsma, K.J. (1997) Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. *Ann. Missouri. Bot. Gard.* 84:1-49.

Soltis, E.D. & Soltis, P.S. (2000) Contributions of plant molecular systematics to studies of molecular evolution. *Plant Mol. Biol.* 42:45-75.

Soltis, P.S., Soltis, E.D. & Chase, M.W. (1999a) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 402:402-404.

Soltis, P.S., Soltis, D.E., Wolf, P.G., Nickrendt, D.L., Chaw, S-M., & Chapman, R.L. (1999b) The phylogeny of land plants inferred from 18S rDNA sequences: Pushing the limits of rDNA signal? *Mol. Biol. Evol.* 16:1774-1784.

Stiller, J.W. & Hall, B.D. (1997) The origin of red algae: Implications for plastid evolution. *Proc. Natl. Acad. Sci. USA*. 94:4520-4525.

Stiller, J.W. & Hall, B.D. (1998) Sequences of the largest subunit of RNA polymerase II from two red algae and their implications for rhodophyte evolution. *J. Phycol.* 34:857-864.

Stiller, J.W., Duffield, E.C.S. & Hall, B.D. (1998) Amitochondriate amoeba and the evolution of DNA-dependent RNA polymerase II. *Proc. Natl. Acad. Sci. USA*. 95:11769-11774.

Stiller, J.W., Riley, J. & Hall, B.D. (2001) Are red algae plants? A critical evaluation of three key molecular data sets. *J. Mol. Evol.* 52:527-539.

Strimmer, K. & von Haeseler, A. (1996) Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13: 964-969.

Thomas, B.A. & Spicer, R.A. (1987) The evolution and Palaeobiology of land plants. Ed. Dudley, T.R., Dioscorides Press, U.S.A.

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

Wang, S.X, Hunter, W. & Plant, A. (2000) Isolation and purification of functional total RNA from woody branches and needles of sitka and white spruce. *BioTechniques*. 28:292-296.

Weier, T.E., Stocking, C.R., Barbour, M.G. & Rost T.L. (1982) Botany: An introduction to plant biology. 6th Edition. Chapter 23, 27-30. John Wiley & Sons Inc. U.S.A.

Winter, K-U., Becker, A., Munster, T., Kim, J.T., Saedler, H. & Thiessen, G. (1999) MADS-box genes reveal that gnetophytes are more closely related to conifers than to flowering plants. *Proc. Natl. Acad. Sci. USA*. 96:7342-7347.

Wolfe, S.L. (1993) Molecular and Cellular Biology. Wadsworth Publishing Company. USA. p.580.

Zillig, W., Klenk, H-P., Palm, P., Puhler, G., Gropp, F., Garrett, R.A. & Leffers, H. (1989) The phylogenetic relations of DNA-dependent RNA polymerases of archaebacteria, eukaryotes, and eubacteria. *Can. J. Microbiol.* 35:73-80.

Appendix 1

Table 6. A complete list of internal primers used to sequence the RPB1 gene.

Name	Motif	Sequence	Tm^a (°C)
LCintf1.1	RKKADEQE	5' CACGCAAGAAAGCCGATGAGCAAG 3'	62
RPB1intf	ARPDWMI	5' GCICGICCIGAYTGGATGAT 3'	50
Cycintr5	HIATYFD	3' CAAAATARGTIGCAATGTG 5'	46
Cycintf2	AKEGRI	5' GCCAAGGAGGGTCGGAT 3'	56
RPB1intr4	AKEGRI	3' ATCCKRCCYTCTTTRGC 5'	52
NWintr	GNLMGKR	3' CGYTTYCCCATYAIRTTYCC 5'	53
RPB1intf2	YIIREDGQ	5' TAYATHATHMGIGAIGAYGGICAR 3'	47
Psiintf3	YKVERHL	5' ATAAGGTGGARAGRCATYTG 3'	52
Cycintr4	YKVERHL	3' CAAATGCCTTTCCACCTTAT 5'	51
Psiintr3	MGHKIRIM	3' TGATACGAATCTTGTGTCCCATAA 5'	60
RPB1intf3.1	AEVLELMMV	5' GCAGAAGTCTTGGARYTGATGATGG 3'	61
RPB1intr3.2	VMGIVQ	3' GSACAATACCCATAACMG 5'	51
Psiintr4	WTGKQ	3' TTGCTTSCCAGTCCA 5'	46
Cycintf3	WTGKQ	5' TGGACTGGSAAGCAA 3'	46
Zamintf2	RIEKEKV	5' CGTATTGAGAAAGAGAAGGTT 3'	52
RPB1intf4	WEEVGPDA	5' TGGGAAGAGGTKGGTCCAGATGC 3'	63
LCintr3	GHTQWLV	3' CAAGCCAYTGTGTRTGMCCY 5'	57
RPB1intr2	QWLVNYW	3' CCARTAYTTIACIAGCCAYTG 5'	44
RPB1intf5	AMGGREG	5' GCWARGGGWGGTAGRGAAGGK 3'	60
LCintr2	EGLIDTAV	3' CACAGCAGTRTCTATCARMCCTTC 5'	60
RPB1intr	YGEDGM	3' CATICCRTCYTCICCRTA 5'	44
LCintr1.1	EFENVYRY	3' GCATCAAACACATTCTGAAACTC 5'	55

a, Tm = melting temperature