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.

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

Université d'Ottawa · University of Ottawa

THE EVOLUTION OF LAND PLANTS INFERRED USING DOMAINS D-F OF THE LARGEST SUBUNIT OF RNA POLYMERASE II

by

SHEHRE-BANOO MALIK

Thesis submitted to the
School of Graduate Studies and Research
University of Ottawa
in partial fulfilment of the
MSc degree
at the
Ottawa-Carleton Institute of Biology

Ottawa, Ontario, Canada

12th January 2000

ACKNOWLEDGEMENTS:

I would like to thank Dr. Guy Drouin for giving me the opportunity and major guidance for this project. Guy, Dr. Donal Hickey, Dr. Lynn Gillespie and Dr. George Carmody -- members of my supervisory committee -- are thanked for all of their constructive advice throughout this project. Many thanks to Michael Ell for taking the time to guide me around. Dr. Andrew J. Roger, Dr. Sandie L. Baldauf, Dr. Hans-Peter Klenk, Dr. Ben D. Hall, Betty McConaughey, Dr. John W. Stiller, Tim Chipman, Dr. Peter G. Foster, Dr. Selvi Subramanian and Dr. John M. Logsdon Jr. are thanked for helpful discussions and practical advice regarding methodology. J. Stiller and J. Logsdon also kindly shared the RNA polymerase alignments from their latest publications. The *Giardia* RPB1 sequence is from H.-P. Klenk. Thanks to Marc Carrier for help with sequencing and for his dedicated work. Thanks to Mehrdad Babaei for providing two *Ephedra viridis* RPB1 (D-F) clones. Many thanks to Guy for the Southern analysis of *Magnolia* and *Nymphaea*. Cathy, Ed, Lia, Patrick, Gitanjali and Dave are thanked for their guidance and friendship. Special thanks also go to Louise Cossette, Lise Maisonneuve, Caroline Pharand, Annie Robillard and the department's technical staff for always being pleasantly helpful. Dr. David Currie, Dr. Jim Fenwick, Dr. Linda Bonen and Dr. Scott Findlay are thanked for their constructive remarks. I would like to thank each of my labmates for their patience, friendship, teamwork and for the pleasure of their company.

I am indebted to my mother and grandmother, for always stressing the importance of a good education and a decent life, and to my young brothers for filling us with hope and youth. I have not achieved anything without their endless love, perseverance and sense of humor.

TABLE OF CONTENTS

ABSTRACT

The aim of this project has been to develop the use of the nucleotide and amino acid sequences of the largest subunit of RNA polymerase II (RPB1) to study the evolutionary relationships of diverse land plants. 1.3 kb of this gene, which includes 900bp of coding sequence corresponding to regions D-F of the protein, have been cloned and sequenced from genomic DNA of seven plant species, one from each of seven divisions. 900 bp of cDNA sequence of regions D-F of RPB1 from *Zea mays* revealed identity with the genomic sequence. Phylogenetic analysis of these nucleotide and corresponding amino acid sequences by parsimony, neighbor-joining and maximum likelihood methods revealed inconsistent results. The maximum-likelihood method presented the best resolved trees. It is unlikely that the inconsistencies are the result of compositional biases, multiple substitutions, or unequal substitution rates. Analysis of amino acid sequences of regions D-F and F-H of RPB1 in animals, plants and fungi showed that animals are more closely related to plants than to fungi, contrary to recent literature. In comparing regions D-F of the largest subunits of RNA polymerases I, II and III (RPA1, RPB1 and RPC1) of all eukaryotes, though, the evolutionary relationships of the major eukaryotic kingdoms were not resolved but RPA1 was found to be more closely related to RPB1 than to RPC1. Since analyses in the literature of longer RPB1 sequences from fewer taxa did clearly resolve relationships of major eukaryotic lineages, this study shows that region D-F of this gene is too short to be informative. Due to the conserved nature of this partial sequence, it appears that a longer sequence from this gene with more synapomorphic sites is required to better assess the evolutionary relationships of plants.

List of Figures and Appendices

LIST OF TABLES

INTRODUCTION

The phylogenetic relationships of the major groups of land plants, as well as of the basal clades of angiosperms, continually confound plant systematists. The rapid, ancient divergence of these taxa and the lack of sufficient extant intermediate taxa are major contributors to this systematic problem. While phylogenetic analyses of morphological features include data from the fossil record, these analyses lack the sensitivity afforded by the phylogenetic analysis of molecular data. Integration of both types of datasets is helpful, but they also conflict in some cases.

While both morphological and molecular analyses have their merits and pitfalls, molecules offer more characters with discretely different states, and increasingly stringent analyses of molecular data are being developed. Unlike morphological datasets, molecular datasets do not yet include data from fossils. Fossilization is rare, though, and even fossils do not yet provide enough intermediates to extant plant lineages (Ridley, 1996). Molecular datasets do not all evolve at the same rate, and sometimes they include enough compositional bias to misgroup taxa (Li and Graur, 1991; Foster and Hickey, 1999). For example, ribosomal RNA genes are affected more by compositional bias and heterogeneous rates of evolution than protein-coding genes are and such genes also evolve at different rates whether they are encoded in the nucleus, chloroplasts, or mitochondria (Hasegawa and Hashimoto, 1993; Li and Graur, 1991). In plants, each of these genomes is subject to different types or degrees of rearrangements. Even within these contexts, different genes also evolve at different rates, so it is important to find molecular phylogenetic markers which accurately represent the phylogeny of the taxa in question, with significant statistical support. Thus, it is useful to seek several independent datasets to corroborate with previous phylogenetic inferences and make our understanding of land plant phylogeny more precise.

Phylogenetic analysis of morphological data in plants falls into several categories: classification according to shared biochemical and cellular properties of extant green algae and plants, comparisons of features of extant and fossilized pollen, and comparisons based on developmental features and ultrastructure of extant plants and fossils when possible (Bremer, 1985; Crane *et al.* 1995; Kenrick and Crane, 1997; Mishler and Churchill, 1985). It was difficult to perform early such analyses due to a lack of clearly defined analytical criteria and raw data (Meacham, 1994; Takhtadzhian, 1997). In the past two decades, however, the accumulation of more fossil evidence and technological advances in the computation of phylogenetic relationships have clarified the relationships of land plants and angiosperms more than ever (Bateman *et al.* 1998; Crane *et al.* 1995; Donoghue, 1994; Kenrick and Crane, 1997). Based upon this information, it is clear that land plants arose from green algae, and that the earliest land plants were bryophytes, from which vascular plants arose as a monophyletic group (Bateman *et al.* 1998 and references therein; Donoghue, 1994). This data also makes it clear that angiosperms are monophyletic and are composed of Eudicots, which have triapturate pollen, and the Monocots and Magnoliid dicots, which have monocolpate pollen (Bateman *et al.* 1998, Crane *et al.* 1995; Taylor and Hickey, 1996; Loconte, 1996). However, the relationships amongst the early land plants are unclear, as are the relationships of the basal clades of angiosperms (Bateman *et al.* 1998; Donoghue, 1994). The relationships are resolved differently depending upon the combinations of characters used, the equal or uneven distribution of taxa sampled, and whether maximum parsimony, maximum likelihood, or neighbour-joining methods of phylogenetic inference are used (Bateman *et al.* 1998; Taylor and Hickey, 1996; Takhtadzhian, 1997; Meacham, 1994; Donoghue, 1994).

This data supports two opposing hypotheses for the origin of angiosperms. The first is that of a herbaceous origin of angiosperms, meaning that the Nymphaeales are the most basal angiosperm clade and monocots and eudicots such as *Arabidopsis* are more derived. The second hypothesis is that woody Magnoliids are most basal amongst angiosperms, and that they share some monocolpate pollen features with gymnosperms and others with monocots. Clearly, the abundance of homoplasy and the shortage of synapomorphies when comparing such diverse taxa as angiosperms and other land plants is a problem, as with any other anciently diversified group. This problem is hampered by the unequal rates of evolution of phenotypic characters in these groups, gaps in the fossil record and unbalanced and unequal sampling of extant taxa. A proper understanding of plant phylogeny requires that these gaps are filled.

Similar problems of homoplasy and unequal rates of evolution are faced by molecular systematists. The Siluro-Devonian radiation of land plants, approximately 480-450 million years ago (MYA), is estimated to have occurred over 35 - 100 million years, which is very rapid in comparison to the amount of time that has passed between then and now (Bateman *et al.* 1998; Kenrick and Crane, 1997; Stewart and Rothwell, 1993). Such relationships are difficult to measure since molecules (or characters) evolving fast enough to bear a record of the rapid radiation have been saturated with homoplasy by now, and molecules or characters evolving slowly enough to be less homoplastic likely evolved too slowly to be informative of the rapid radiation. An assemblage of molecular data could overcome problems of unequal rates of evolution and homoplasy, since different molecular markers and corresponding models of evolution could be applied to reconstruct different parts of the plant phylogenetic tree (Ritland and Eckenwalder, 1992). Donoghue (1994) pointed out that from 1989-1991, only about 0.6% of phylogenetic analyses considered molecular

data for analysis of the relations of the major groups of vascular plants, land plants or seed plants. Thus, it is imperative to assemble an integrated approach to the molecular systematics of plants, incorporating data from multiple loci and from large-scale genome characteristics such as inversions, insertions/deletions, and introns (Bateman *et al.* 1998; Qiu and Palmer, 1999).

Several molecular datasets from each of the three genomes (nuclear, chloroplast and mitochondrial) have been examined to try to solve the relationships of plants, but conflicts remain. The bulk of molecular phylogenetic analysis of plants lies in the examination of nuclear 18S ribosomal RNA genes from 223 plants (Soltis *et al*, 1997) and of the chloroplast-encoded gene for the largest subunit of ribulose-1,5-bisphosphate carboxylase (*rbcL*) in 499 plants (Chase *et al*, 1993), almost entirely from angiosperm species. Chaw *et al* (1997) examined 18S rRNA in 65 gymnosperms and angiosperms. Nuclear 26S rRNA was examined in 15 angiosperms by Kuzoff *et al* (1998). Hasebe *et al* (1992, 1993) examined the phylogeny of 10-15 seed plants and ferns using *rbc*L. Mitochondrial introns and RNA editing have been examined in three genes from 28 plant species spanning the range of all plants (Malek *et al*, 1996; Qiu *et al*, 1998; Qiu and Palmer, 1999; Cho *et al*, 1998). Multigene families of nuclear genes encoding actin, phosphoenol pyruvate carboxylase (PEPC), floral homeotic MADS-box genes and phytochromes have also been examined in up to 20 diverse land plant taxa (Moniz de Sá and Drouin, 1996 and unpublished data; An *et al.* 1999; Gehrig *et al*, 1998; Purugganan, 1997; Kolukisaoglu *et al.* 1995; Donoghue and Matthews, 1998). The second-largest subunit of RNA polymerase II (RPB2), a single-copy ubiquitous nuclear-encoded protein-coding gene, was also recently used to infer the phylogeny of 12 green plants by Denton *et al* (1998). These are some of the most notable recent examples of molecular markers

applied to resolve the deep phylogenetic relationships amongst green plants and amongst the basal clades of angiosperms.

While molecular data support the origin of land plants from Charophycean algae and the monophyly of land plants, of vascular plants, of Gnetales and of angiosperms, some ambiguities still lie in the early relations of land plants, amongst gymnosperm taxa, and amongst the basal angiosperm clades. The analyses which included the broadest range of taxa were made with molecular markers which evolve slowly, such that they are more suited to studying more recent evolutionary events amongst plants. Other molecular analyses (actin, PEPC, MADS-box genes, phytochrome genes, RPB2) lack the breadth of taxa included in the rRNA and *rbcL* analyses. It is possible that orthologous members of the multigene families may clarify the phylogeny of subsets of plants, by rooting such genes from the ingroup with their latest paralogue in the outgroup (Donoghue and Matthews, 1998; Brown and Doolittle, 1995). In comparison to angiosperms, gymnosperms are underrepresented by many of these analyses, but not as underrepresented as early nonvascular and vascular plants are. While the absence of certain mitochondrial introns supports that liverworts such as *Marchantia* are the earliest land plants (Qiu *et al*, 1998) the authors also emphasize the need to combine data from multiple genes subject to different evolutionary pressures to reduce the influence of homoplasy (Qiu and Palmer, 1999). Also, it is important that all of these datasets be analysed in a manner that minimizes the effects of GC content bias, saturation of synonymous substitutions, and heterogeneous rates of evolution amongst the ingroups, which can result in artefactual phylogenies.

The largest datasets amongst these analyses are those of *rbcL* and 18S rRNA. The *rbcL* dataset supports monophyly of Gnetales, of angiosperms and of monocots and a herbaceous origin

of angiosperms, with monocots either basal to or amongst the paleoherbs (such as *Nymphaea)*, depending on the methods of phylogenetic tree inference used (Chase *et al*, 1993). The 18S rRNA dataset also supports the monophyly of Gnetales, of angiosperms and of monocots, but does not clearly resolve any of the relationships of the major angiosperm clades (Soltis *et al*, 1997). In parsimony analyses of rRNA data, a Magnoliid origin of angiosperms is only one step less likely than a herbaceous origin is (Nickrent and Soltis, 1995).

Only two of the datasets briefly described above include large protein-coding genes. The RPB2 and phytochrome datasets are 3564 bp and 3264 bp in length respectively, including the third base of codons (Denton *et al*, 1998; Donoghue and Matthews, 1998). Protein-coding genes are phylogenetically useful since most of the GC content bias, if any, is absorbed in the synonymous sites of codons, which can be removed from the datasets used for phylogenetic inference. Also, alignments made at the amino acid level and then fitted to corresponding nucleotide sequences can be made with less ambiguity than with nucleotides alone when studying distantly related taxa (Sidow and Wilson, 1990). Large protein-coding genes are also phylogenetically useful since they provide more informative sites for analysis. To date, analyses of full-length datasets of these genes which include diverse gymnosperm, vascular and nonvascular plant taxa have not been shown. The largest subunit of RNA polymerase II, RPB1, has approximately 5 kb of protein-coding sequence in plants (Nawrath *et al*, 1990). The only full-length RPB1 gene available from plants is from *Arabidopsis*. Since RPB1 is a key component of the RNA polymerase II holoenzyme in eukaryotes which transcribes messenger RNA and small nuclear RNAs, it would be useful not only from the phylogenetic point of view but also from a biochemist's point of view to further characterize this gene and protein in land plants.

RPB1 has recently been used to infer phylogenetic relationships of deeply diverging groups of eukaryotes with better statistical support than with data from rRNA, elongation factor (EF-1α, EF-2) and heat shock protein (Hsp70) sequences (Hirt *et al*, 1999; Stiller *et al*, 1998; Stiller and Hall, 1997 and 1998). The Archaeal homologue to the largest subunit of eukaryotic RNA polymerases I, II and III, called the A' subunit, also provided the data for a better-resolved inference of Archaeal phylogenetic relationships than the rRNA dataset had (Puhler *et al*, 1989a; Iwabe *et al*, 1991; Klenk and Zillig, 1994). Partial RPB1 sequences have also been used recently to infer the phylogeny of invertebrates, arthropods, nematodes, and *Leishmania* species (Sidow and Thomas, 1994; Regier and Schultz, 1997; Baldwin *et al*, 1997; Croan and Ellis, 1996). These data show that RPB1 is useful for studying deep-level and later levels of eukaryotic phylogeny. RPB1 is ubiquitous and usually found as a single-copy gene amongst eukaryotes. An RPB1 dataset can thus offer a phylogeny of orthologous molecules, encoded in the nuclear genome, with more informative sites than previous plant molecular datasets and less compositional bias than incurred by non-protein-coding sequences such as rRNA (Sidow and Wilson, 1990).

While the paucity of available plant RPB1 sequences and the homoplasy accumulated in other such ancient characters may discourage the use of RPB1 for inferring land plant phylogeny, the results of other phylogenetic analyses of RPB1 and its homologues are very encouraging. The largest subunits of eukaryotic RNA polymerases I, II and III (RPA1, RPB1 and RPC1), archaeal A' and A" RNA polymerases, and bacterial β and β' RNA polymerases are collectively known as type A RNA polymerases due to their shared homology (Young, 1991). Their protein sequences share eight conserved domains, termed A-H (Young, 1991; Palenik, 1992). RPB1 is also characterized by carboxy-terminal repeats (YSPTSPX), considered as conserved domain "I", which are correlated

with the presence of spliceosomal introns in eukaryotic genes (Stiller and Hall, 1998). These conserved domains simplify the alignment of RNA polymerase genes from highly divergent taxa. The conserved domains are also very suitable for designing degenerate PCR primers for amplifying fragments of RPB1 from all eukaryotes. Sequences interspersed between domains A-I which are less conserved and are difficult to unambiguously align amongst deeply diverging protist lineages can, however, provide informative phylogenetic data when examining more closely related species.

Domains D and F of RPB1 provide a useful starting point for the examination of RPB1 in land plants. The conserved motif "PYNADFDGDEMN" in domain D is involved with domain G in elongation and contains a DNA binding element of RPB1 which is found in all type A RNA polymerase homologues. The aspartic acid residues bind Mg^{++}. The nucleotide sequence for this motif in *Arabidopsis* RPB1 also lacks any introns, repeats or sixfold degenerate codons, is rich in twofold degenerate codons, and is long enough to provide an RNA polymerase-specific PCR primer with a stringent annealing temperature (i.e., high enough to avoid non-specific priming). Domain F, which is involved in elongation, contains a motif "VGQQNVEG" which represents the alpha-amanitin sensitive region which is a characteristic of RPB1. This motif, as stated, is not as highly conserved amongst all eukaryotes as is the domain D motif, but is highly conserved amongst crown eukaryotic lineages, and is the consensus of *Arabidopsis thaliana* and *Glycine max* plant RPB1 sequence for that domain. In these plants, the nucleotide sequence for this motif also lacks introns or sixfold degenerate codons and contains a few twofold degenerate codons, making it a stringent and useful candidate for a PCR primer as well. While these motifs only span 900 bp of coding sequence for RPB1 in *Arabidopsis*, that region should be easily obtained by PCR and provide a valuable starting point for further characterization of RPB1 in land plants.

On the basis of these characteristics of RPB1 and the successful use of complete and partial sequences of this gene in addressing a variety of phylogenetic questions, I hypothesize that it will be a useful tool for studying land plant phylogeny. In addition to this, characterization of RPB1 from more plant taxa will eventually lead to a better understanding of its fundamental biochemical properties in transcription.

MATERIALS AND METHODS

Plant Material:

Plant material for this study was collected from greenhouses at Carleton University and the University of Ottawa, the Rideau River and from a park in Cumberland, with the kind and generous assistance of Hughette Allard, Ed Bruggink and Elizabeth Thompson. Species names are as indicated in Table 1.

Plant DNA and RNA purification:

Plant total genomic DNA was prepared using the CTAB method of Doyle and Doyle (1990) and using the Qiagen Plant DNeasy Maxi and Mini Kits exactly as directed by the manufacturer. Two modifications to the CTAB method were that the extraction buffer included 2% polyvinylpyrrolidone and that if a thick, creamy interface persisted after the chloroform extractions, then a phenol-chloroform extraction was performed. Genomic DNA preparations from *Ephedra tweediana*, *Gnetum gnemon*, and *Welwitschia mirabilis* were kindly provided by Dr. Lynn Gillespie. CsCl-purified genomic DNA for *Zea mays* was obtained from Dr. Guy Drouin. These gifts of DNA as well as DNA prepared using the CTAB method from all of the other plant species included in this project were used as templates for PCR amplification of RPB1. While suitable as PCR templates, some of these DNA preparations still contained polysaccharides or secondary metabolites which interfered with restriction enzyme digestion, hindering Southern blot analysis. In such cases, either existing total DNA preparations were further purified using the Qiagen Plant DNeasy kit, or the same kit was used to prepare more DNA from fresh tissue.

Table 1: Plants from which RPB1 D-F were PCR amplified for this analysis.

Taxa	***Species Name***	***Common Name***
Coniferophyta	*Araucaria heterophylla*	Norfolk Island pine
Cycadophyta	*Cycas revoluta*	cycad
Gnetales, Ephedraceae	*Ephedra viridis*	Mormon tea
Gnetales, Ephedraceae	*Ephedra tweediana*	Mormon tea
Equisetophyta	*Equisetum hyemales*	horsetail
Gingkophyta	*Ginkgo biloba*	gingko
Gnetales, Gnetaceae	*Gnetum gnemon*	gnetum
Magnoliales	*Magnolia soulangeana*	magnolia
Marchantiopsida	*Marchantia polymorpha*	liverwort
Nymphaeaceae	*Nymphaea odorata*	water lily
Filicophyta	*Osmunda claytonea*	fern
Psilotophyta	*Psilotum nudum*	whisk fern
Gnetales, Welwitschiaceae	*Welwitschia mirabilis*	welwitschia
Liliopsida (monocots)	*Zea mays*	maize/corn

Plant total RNA was prepared according to the CTAB method of Chang *et al* (1993) and using the Qiagen Plant RNeasy Kit as directed by the manufacturer. RNA prepared using the CTAB method was successfully used as a template for RT-PCR of RPB1 from maize. RT-PCR was unsuccessful using RNA prepared using the CTAB method from *Cycas revoluta, Equisetum hyemales, Marchantia polymorpha* and *Psilotum nudum.* This, too, was inhibited by secondary metabolites that could not be separated from the RNA. The Qiagen RNeasy kit was later used to successfully further purify total RNA from *E. hyemales* that had been prepared three years previously by the CTAB method. The Qiagen RNeasy kit was then used to successfully prepare total RNA from fresh tissue of *Magnolia soulangeana, Nymphaea odorata, Osmunda claytonea, Ephedra viridis* and *Zea mays* (Pioneer variety 3953). RNA from *Araucaria heterophylla* and *Ginkgo biloba* could not be obtained using this method, though, so the CTAB method was used to obtain RNA from these taxa.

Amplification of RPB1:

By aligning RPB1 sequences from animals, fungi, *Arabidopsis* and soybean (*Glycine max)* using ClustalV (Higgins, 1994), conserved regions (A-I) were identified (Figure1), and degenerate oligonucleotide primers were designed, with reference to conserved amino acid motifs identified by Sidow and Thomas (1994). The primers correspond to the ECPGHFGH motif in domain A (5' GAI TGY CCI GGI CAY TTY GG 3' forward), the PYNADFDGDEMN motif in domain D (5' CCI TAY AAY GCI GAY TTY GAY GGI GAY GAR ATG AA 3' forward) and the VGQQNVEG motif in domain F (5' CC YTC IAC RTT YTG YTG ICC IAC 3' reverse). Table 2 represents sections of the RPB1 amino acid alignment from which D and F PCR primers were designed, including parts of

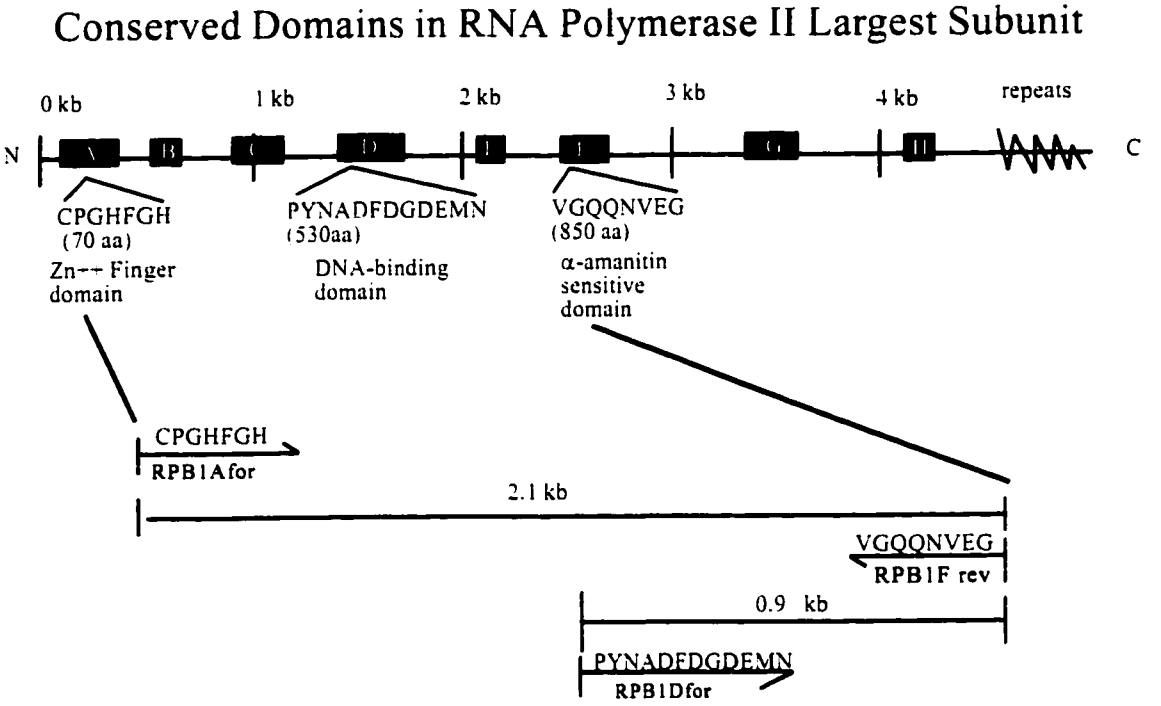


Figure 1: Schematic diagram of the coding region of RPB1, with the blocks of conserved amino acids highlighted to identify the positions of corresponding PCR primers. Zigzag lines indicate the carboxy terminal repeats, while vertical lines represent the size of the protein-coding nucleotide sequence.

Table 2: Alignments corresponding to domains D and F of RPB1, including chloroplast (cp), *E. coli*, archaeal and RPA1 and RPC1 homologues. Corresponding species names are given in Table 3. Domain D is specific to type A RNA polymerases. Domain F is characterized by an α-amanitin sensitive domain only in RPB1,and not in other homologues. Amino acid residues identical to *Arabidopsis* sequence are indicated by a dot.

Species	*Domain D*	*Domain F*
Arab.th-II	**PYNADFDGDEMN**	**VGQQNVEG**
Sac.cer-II	.....	....S...
Sch.pom-II	.....	....I...
Try.bru-II	.....	......A.
Cae.ele-II	.....	.....
Dro.mel-II	.....	.....
Mus.mus-II	.....	.....
Sac.cer-I	A.....	L...AL..
Try.bru-I	SF...T.....	L...LFD.
Sac.cer-III	.....	....IIS.
Try.bru-III	.....	....T.S.
H.halobium A'	.....	....A.R.
E.coli β'	A.....Q.A	R.LMAKPD
Marchantia cp	GF.....Q.A (β')	R.LMSDPQ (β")

alignments by Puhler *et al* (1989a). This alignment shows that the amino acid sequence encoded by primer D is specific to RNA polymerases, and that the amino acid sequence encoded by primer F is specific to RPB1, and that they would not amplify RNA polymerase genes from chloroplast and mitochondrial genomes.

All primers were synthesized by Cortec DNA Service Laboratories or by Life Technologies. PCR reactions were performed following cycling conditions of: 2 minutes at 94°C, 35 cycles of 40 seconds at 94°C, 1 minute at 55°C and 3 minutes at 74°C, followed by 10 minutes at 74°C. 200 ng of plant genomic DNA was used in PCR amplifications consisting of 2.5 - 5.5 mM $MgCl_2$, 0.75 μM each primer, 200 μM each dNTP, 0.025 U/μl *Taq* DNA polymerase, 1 X PCR buffer (supplied by the manufacturer of the *Taq* polymerase), and an additional buffer described by Ponce and Michol (1992) consisting of 30 μM tricine pH 8.4, 5 μM 2-mercaptoethanol, 10 μg/ml BSA and 0.05% Tween-20. PCR reagents were obtained from Promega, Pharmacia, Boehringer-Mannheim, and BDH.

Positive, negative and single primer controls were included in each batch of PCR amplifications. A clone of *Arabidopsis* RPB1 (pOSA2, obtained from R. Larkin and T. Guilfoyle) was used as the template for the positive control to show that the amplification conditions worked. No DNA template was included in the negative control, intended to indicate if any contaminants were present in the reagents. Single primer controls were made to indicate if any of the PCR products were artifacts which resulted from nonspecific binding of a single primer.

Southern blots of D-F PCR amplifications from genomic DNA of 13 plants were probed with the *Arabidopsis* RPB1 clone. This revealed a 1.3 kb putative RPB1 PCR product from each of the

13 plants. This corresponded to the 900 bp of coding sequence in addition to three introns found between my domain D and F RPB1 primers in *Arabidopsis*.

First-strand cDNA was prepared from total maize RNA using the Amersham First-Strand cDNA Synthesis Kit, according to the manufacturer's instructions, priming with random hexanucleotides. One-tenth of the cDNA reaction was then used directly as the template for an RPB1 PCR amplification, as described above.

Cloning:

PCR products of the expected size were isolated from agarose gels and purified using the GeneClean II Kit (Bio 101). Purified products from maize genomic templates were blunt-ended with Klenow fragment and T4 polynucleotide kinase purified, and ligated into the dephosphorylated *Eco*RV site of pBluescript II SK+ (Stratagene) with T4 DNA ligase at 10°C overnight. This was inefficient, and cloning attempts with the Stratagene pCRScript SK+ blunt-ended cloning kit were not fruitful, so all other RPB1 PCR products were cloned using the TOPO TA Cloning Kit (Invitrogen) into pCR2.1 according to the manufacturer's instructions. Ligations were transformed into *E. coli* strain DH11S by heat shock into RbCl-competent cells or using the Transformaid Kit according to the manufacturer's instructions (MBI Fermentas), or heat-shock transformed into OneShot *E. coli* cells as directed (Invitrogen). Transformants were grown using Luria-Bertani agar and broth under ampicillin selection, with the addition of IPTG and X-gal to the agar to facilitate blue/white colour selection of transformants (Sambrook *et al* 1989). Enzymes were from New England Biolabs, Pharmacia, Stratagene and Boehringer-Mannheim. LB agar and broth components were from Difco and BDH.

Putative clones were identified by colony hybridization or according to the size of their inserts. Insert size was found either by restriction analysis of the plasmids or by PCR amplification of the vector's multiple cloning site using M13 forward and reverse sequencing primers (adapted from Sandhu *et al*, 1989).

Plasmids were prepared by the alkaline lysis method of Sambrook *et al* (1989), or with the Eclipse Mini Kit or QIAprep Spin Miniprep Kit, according to the manufacturer's directions (Gordon Technologies, Qiagen), with the following modifications: After the addition of lysis solution, the cells were incubated at room temperature to lyse for 3-5 minutes prior to the addition of the high-salt buffer. Upon addition of the salt, the preparations were microcentrifuged for 7-10 minutes at 13 krpm to pellet all of the cellular debris.

In order to release their insert, clones in the pCR2.1 vector were analysed by digestion with *Eco*RI, those in pBluescript SK+ were digested with *Eco*RI and *Bam*HI and clones in pCRScript SK+ were digested with *Eco*RI and *Sac*I. 2 ml of the plasmid prep (125 -400 ng) were completely digested, as directed, in a final volume of 15 µl and then visualized on a 0.7% agarose, 1 X TBE gel run at 10 V/cm.

Colony blots were performed as described in the Hybond-N membrane instruction booklet (Amersham) and probed with a homologous probe prepared using P^{32}-labelled dCTP using the Pharmacia T7 QuickPrime Kit, as described by the manufacturer. Membranes were sandwiched between two pieces of Whatman 3MM paper and hybridized at 42°C in 50% formamide, 5 x SSC, 0.5% SDS, 50 µg/ml denatured sheared salmon sperm DNA, 5 x Denhardt's solution. The membranes were washed with gentle agitation three times at room temperature in 1 l of 2 x SSC,

0.1% SDS for five minutes, then washed twice for 15 minutes at 65°C in 1 l of 0.1 x SSC, 0.5% SDS.

PCR screening was performed by using a sterile toothpick to inoculate a white colony into a 10 µl PCR mixture and then to inoculate a patch on an LB agar plate. The PCR mixture consisted of 150 µM each dNTP, 1 µM each primer (M13 forward and M13 reverse, from Invitrogen), 0.025 U/ml *Taq* DNA polymerase and 1 x the PCR buffer containing 1.5 mM $MgCl_2$ supplied with it (Pharmacia or Boehringer-Mannheim). Cycling conditions were: 2 minutes at 94°C followed by 30 cycles of 1 minute at 94°C, 1 minute at 53°C and 1½ minutes at 74°C, and 10 minutes at 74°C. The entire PCR reaction, representing the insert in the cloning vector, was then visualized by agarose gel electrophoresis, as described above.

Single-stranded versions of these RPB1 clones were prepared by a protocol adapted from Garber *et al* (1993). A fresh culture of the clone was inoculated into 2 ml of LB broth with 100 µg/ml ampicillin and 10 mM $MgCl_2$, incubating it at 37°C for two hours, then infecting these cells with F1 helper phage and then increasing the media by 8 ml, followed by an incubation at 37°C for an additional 8 hours before harvesting the phage. To harvest, the culture was centrifuged and the phage precipitated from the supernatant with ¼ volume of 20% PEG, 2.5 M NaCl at 4°C for 1-12 hours. The suspension was centrifuged and the phage pellet resuspended in 1.25 ml TERPS (10 mM Tris, 1 mM EDTA, 40 µg/ml RNase A, 2.5 µg/ml Proteinase K, 0.5% SDS) and lysed for 2 hours at 65°C. The phage coat was precipitated with 1/5 volume of 5M potassium acetate, the suspension centrifuged, and the phagemid DNA was precipitated from the supernatant with 0.7 volumes of cold isopropanol. A pellet of single-stranded DNA was washed twice with cold 70% ethanol, dried and resuspended in 50 µl of TE.

Sequencing:

Single- and double-stranded versions of the RPB1 clones were sequenced according to the Sanger-dideoxy method, by primer walking. Initial sequence was obtained manually using the Pharmacia T7 Sequencing Kit and S^{35}-labelled dATP (Amersham), according to the manufacturer's instructions, with the addition of 3 µl DMSO and annealing at 65°C to resolve compressions caused by secondary structure in the sequence. These reactions were run on 0.4 - 0.8 mm thick wedge-shaped denaturing polyacrylamide gels (4% acrylamide, 0.5 x TBE, 8M urea) at 1700 V for 5 and 17 hours using electrophoretic apparatus from Owl Scientific. Sequencing gels were vacuum-dried (BioRad) onto Whatman 3MM paper and exposed to Kodak BioMax MR film for 1 - 5 days.

Further sequence was obtained using the ABI Cycle Sequencing Core Kit with AmpliTaq FS or the ABI BigDye Terminator Cycle Sequencing Kit with AmpliTaq FS (ABI, Perkin-Elmer Corporation) according to the manufacturer's instructions. These automated sequences were read by the University of Ottawa Biotechnology Research Institute. Plant-specific degenerate internal sequencing primers were designed from my alignment (prepared using ClustalV under default settings) of the sequenced ends of each clone. CODEHOP (Rose *et al*, 1998) was used to help design primers. The primer positions correspond to the IIPKQIN motif (5' CCT GAT CAT CCC GAA GCA RAT HAA YHT 3' forward and 5' T GAT CAG GTT GAA GAC CTG CTT NCC NGT CCA 3' reverse) and the GIGDTIAD motif (5' GGI ATI GGI GAY ACI ATI GCI GA 3' forward and 5' TC IGC IAT IGT RTC ICC IAT ICC 3' reverse).

Northern and Southern Hybridizations:

Northern blots of 20 μg of total RNA were prepared using the Ambion NorthernMax kit as directed, with the following modifications. A 0.8% SeaKem agarose formaldehyde gel, prepared as indicated by Sambrook *et al* (1989) was used to fractionate the RNA at 2V/cm for 18 hours. The RNA was then capillary transferred using alkaline 5 x SSC for 5 hours, and the membrane was hybridized and washed as directed (Ambion).

Southern blots were prepared as indicated by Sambrook *et al* (1989), with the following adaptations. 10 μg of total DNA were completely digested singly with *Eco*RI and *Bam*HI or *Eco*RI, *Pst*I and *Xba*I under standard conditions and electrophoresed on 0.8% SeaKem agarose gels with 1 X TAE buffer at 45 V for 19 hours, then capillary transferred to nylon membranes overnight using 20 x SSC as the transfer medium.

Both Northerns and genomic Southerns were hybridized with their corresponding RPB1 clones. For the Southern blot of PCR products, 10 μl of each PCR was run in a 0.7% agarose gel in 1 x TBE at 10 V/cm for an hour. RNA or DNA was cross-linked to the damp nylon membranes by exposure to ultraviolet light.

Membranes were pre-hybridized for 25 minutes to 2 hours at 42°C in 25 ml of 50% formamide, 5 x SSC, 0.5% SDS, 50 μg/ml denatured sheared salmon sperm DNA, 5 x Denhardt's solution. Hybridization was in 10 ml of this solution at 42°C for 18 - 24 hours. Calf thymus DNA was substituted for salmon sperm DNA in the hybridization solution for the genomic Southern blots. Background hybridization was also avoided by removing unincorporated nucleotides from the probe using Pharmacia ProbeQuant columns (according to the manufacturer's instructions) and by sandwiching the membranes between two pieces of Whatman 3MM paper during the hybridization.

Northern and genomic Southern blots were probed with a homologous probe prepared using P^{32}-labelled dCTP using the Pharmacia T7 QuickPrime Kit according to the manufacturer's instructions.

Membranes were then washed twice at room temperature in 1 l of 2 x SSC, 0.1% SDS for 10 minutes, then washed twice for 10 minutes at 65°C in 1 l of 2 x SSC, 0.1% SDS. The Southern blot of RPB1 PCR products was probed with a heterologous probe (pOSA2, a clone of RPB1 from *Arabidopsis*), prepared and hybridized as described above, but only washed twice at room temperature in 1 l of 2 x SSC, 0.1% SDS for fifteen minutes. The membranes were exposed to Kodak Biomax MS film with Kodak Biomax Transcreen HE intensifying screens at -80°C for 2-7 days.

Sequence Analysis:

Plant RPB1 D-F nucleotide sequences were edited using GDE and Sequedit v.0.912 (Drouin *et al*, 1999). Additional eukaryotic RPA1, RPB1 and RPC1 nucleotide and amino acid sequences were obtained from Genbank using BLAST (Altschul *et al*, 1990 and 1997) and Entrez, except for the *Giardia lamblia* RPB1 DNA sequence which was obtained from Hans-Peter Klenk. Table 3 provides full identification of the sequences used in alignments in this report. Peter Foster's "pgb_cds" PERL script was used to isolate the coding region of these nucleotide sequences. The amino acid alignment of animal, plant and fungal sequences, from which plant-specific RPB1 PCR primers were designed, was prepared using Clustal V (Higgins, 1994). This was done via GDE version 2.0 (Smith *et al*, 1994), using a floating gap penalty of 90 and a fixed gap penalty of 10 besides the default settings, with further manual adjustments made using GDE. My own sequences were identified as plant RPB1 D-F using basic BLAST searches (Altschul *et al*, 1990 and 1997).

Table 3: Codes and sources for sequences included in phylogenetic analyses of type A RNA polymerases.

Species name	Code	Polymerase	Accession #	Authors
Sulfolobus acidocaldarius	*Sulf.acido*	*A'*	1133407	Puhler *et al*, 1989b
Porphyra yezeonsis	*Por.yez-II*	*RPB1*	U90208	Stiller and Hall, 1997
Bonnemaisonia hamifera	*Bon.ham-II*	*RPB1*	U90209	Stiller and Hall, 1997
Drosophila melanogaster	*Dro.mel-II*	*RPB1*	M27431	Jokerst *et al*, 1989.
Artemia salina	*Art.sal-II*	*RPB1*	U10331	Sidow and Thomas, 1994.
Crassostrea gigas	*Cra.gig-II*	*RPB1*	U10334	Sidow and Thomas, 1994.
Ilyanassa obsoleta	*Ily.obs-II*	*RPB1*	U10338	Sidow and Thomas, 1994.
Helobdella stagnalis	*Hel.sta-II*	*RPB1*	U10336	Sidow and Thomas, 1994.
Homo sapiens	*Hom.sap-II*	*RPB1*	X63564	Wintzerith *et al*, 1992.
Mus musculus	*Mus.mus-II*	*RPB1*	M14101	Ahearn *et al*, 1987.
Caenorhabditis elegans	*Cae.ele-II*	*RPB1*	M29235	Bird and Riddle, 1989.
Spirogyra spp.	*Spirogy-II*	*RPB1*	U90210	Stiller and Hall, 1997.
Gingko biloba	*Gin.bil-II*	*RPB1*		this study
Cycas revoluta	*Cyc.rev-II*	*RPB1*		this study
Magnolia soulangeana	*Mag.sou-II*	*RPB1*		this study
Nymphaea odorata	*Nym.odo-II*	*RPB1*		this study
Zea mays	*Zeamays-II*	*RPB1*		this study
Glycine max	*Gly.maxC-II*	*RPB1*	X52495	Dietrich *et al*, 1990.
Arabidopsis thaliana	*Arab.th-II*	*RPB1*	X52954	Nawrath *et al*, 1990.
Ephedra viridis	*Eph.vir-II*	*RPB1*		this study
Araucaria heterophylla	*Arau.he-II*	*RPB1*		this study
Dictyostelium discoideum	*Dic.dis-II*	*RPB1*	S52651, AF058710	Lam *et al*, 1992; Stiller and Hall, 1998.
Acanthamoeba castellanii	*Aca.cas-II*	*RPB1*	U90211	Stiller and Hall, 1997.
Saccharomyces cerevisiae	*Sac.cer-II*	*RPB1*	M11190	Allison *et al*, 1985.
Schizosaccharomyces pombe	*Sch.pom-II*	*RPB1*	S56564	Azuma *et al*, 1991.
Nosema locustae	*Nos.loc-II*	*RPB1*	AF061288	Hirt *et al*, 1999.
Vairimorpha necatrix	*Vai.nec-II*	*RPB1*	AF060234	Hirt *et al*, 1999.
Plasmodium falciparum	*Pla.fal-II*	*RPB1*	133328	Li *et al*, 1989.

Species name	Code	Polymerase	Accession #	Authors
Mastigamoeba invertens	*Mas.inv-II*	*RPB1*	AF083338	Stiller *et al*, 1998.
Trichomonas vaginalis	*Tri.vag-II*	*RPB1*	U20501	Quon *et al*, 1996.
Trypanosoma brucei	*Try.bru-II*	*RPB1*	133329	Evers *et al*, 1989.
Giardia lamblia (intestinalis)	*Gia.lam-II*	*RPB1*		H.-P. Klenk, unpublished.
Saccharomyces cerevisiae	*Sac.ce-III*	*RPC1*	X03129	Allison *et al*, 1985.
Giardia lamblia (intestinalis)	*Gia.la-III*	*RPC1*	P25202	Lanzendorfer *et al*, 1992.
Trypanosoma brucei	*Try.br-III*	*RPC1*	P08968	Kock *et al*, 1988.
Mus musculus	*Mus.musc-I*	*RPA1*	35134	Seither *et al*, 1997.
Drosophila melanogaster	*Dro.mela-I*	*RPA1*	P91875	Knackmuss *et al*, 1996.
Trypanosoma brucei	*Try.bruc-I*	*RPA1*	P16355	Jess *et al*, 1989.
Saccharomyces cerevisiae	*Sac.cere-I*	*RPA1*	P10964	Memet *et al*, 1988.
Schizosaccharomyces pombe	*Sch.pomb-I*	*RPA1*	P15398	Yamagishi and Nomura, 1988.

RPB1 D-F amino acid sequences including my own were later aligned for phylogenetic analysis using ClustalW 1.71 (Thompson *et al*, 1994) under default settings and then refined manually in GDE, with reference to eukaryotic RPB1 alignments obtained from John Stiller (Stiller and Hall, 1997) and John Logsdon (Hirt *et al*, 1999). The coding nucleotide sequences were aligned manually with the amino acid alignment using GDE and Sequedit v.0.912. To reduce the effect of compositional bias introduced by synonymous substitutions, only the first and second bases of codons were used in the analysis. "M" and "y" were substituted for the first codon position of leucine and arginine codons, to reduce compositional bias. P. Foster's "pgb_subset" PERL script was used to remove the third base of codons in the nucleotide sequence alignment. The alignments were truncated to contain only the region of the RPB1 fragment within my PCR primers for regions D and F. PHYLIP v.3.573c (Felsenstein, 1993) and PUZZLE v.4.0.2 (Strimmer and von Haeseler, 1996 and 1997) were used to infer phylogenetic trees of both the amino acid and nucleotide sequences.

Phylogenies were first inferred from nucleotide and amino acid sequence alignments using the PROTPARS, DNAPARS, PROTDIST, DNADIST, NEIGHBOR, SEQBOOT, and CONSENSE programs available from PHYLIP v.3.573c (Felsenstein, 1993). The jumble option was used wherever possible to avoid bias from the input order of sequences. The input order of bootstrapped datasets was only randomized once. Global rearrangements were allowed. The PROTDIST algorithm was set to use the Dayhoff PAM model of nucleotide substitution, while both the Kimura and maximum likelihood models and remaining default settings were used for DNADIST. Ten replicates were made of the distance analyses, jumbling once using ten different random seed numbers. Both amino acid and nucleotide multiple sequence alignments were subjected to 1000

bootstrap replicates using SEQBOOT, and consensus distance and parsimony trees were obtained using CONSENSE.

Next, PUZZLE v. 4.0.2 (Strimmer and von Haeseler, 1996 and 1997) was used to infer phylogenetic trees using my plant nucleotide and amino acid alignments, in addition to alignments of RPB1 D-F amino acid sequences amongst animals, plants and fungi, or including RPA1, RPB1 and RPC1 D-F of all eukaryotes. This program uses the quartet puzzling and maximum likelihood approaches to phylogenetic inference. The parameters for the plant amino acid and nucleotide analyses included 10000 puzzling steps, exact parameter estimates, quartet sampling, neighbor-joining tree-fitting, the computation of both clock like and non-clock like branch lengths and use of a mixed model of substitution rate heterogeneity which includes an invariable rate and eight gamma rates. The JTT (Jones *et al.*, 1992) and Blosum 62 (Henikoff and Henikoff, 1992) models of amino acid substitution and the Schoeniger-von Haeseler model of nucleotide substitution were used. Nucleotide analyses were made with this program once by selecting all sites for the analysis; next, selecting only the first and second sites of codons (with "m" or "y" in the first position of leucine and arginine codons); and finally, selecting only the second site of codons. The phylogenetic trees inferred using PUZZLE and PHYLIP were drawn with Treeview and saved as ".wmf" (Windows Metafile) format files, which were then imported into Corel Presentations 8 for further labelling.

Nucleotide frequencies of each sequence at the different codon positions were measured using CODONS 1.4 (Lloyd and Sharp, 1992), which was also used to calculate the amino acid composition. Weighted numbers of synonymous (K_s) and non-synonymous (K_a) substitutions were calculated using Li93 (Li, 1993). The proportion of different synonymous (P_s) and nonsynonymous (P_a) sites was calculated using the equation $P = \frac{3}{4}(1-e^{(4/3)K})$. RRTree v.0.6 (Robinson *et. al.*, 1998)

(P_a) sites was calculated using the equation $P = \frac{3}{4}(1-e^{(4/3)K})$. RRTree v.0.6 (Robinson *et. al.*, 1998) was used for relative rate test calculations. MS Excel '97 was used to construct all graphs. All sequence analysis programs were used in the context of Windows98, except for pgb_cds, pgb_subset, Li93, GDE and ClustalV, which were used on a Sun Sparc 4 workstation.

RESULTS

Preliminary Analysis of the Region Between Domains F and H of RPB1:

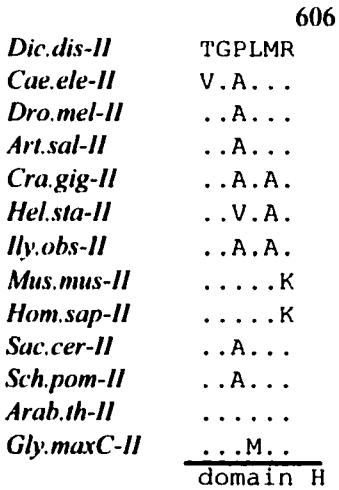
To begin the project, an alignment of published RPB1 amino acid and nucleotide sequences from Animals, Plants and Fungi, with *Dictystelium discoideum*, a slime mold, as an outgroup, was prepared to use to estimate the usefulness of this gene for inferring the phylogeny of land plants and to use as a guide for designing PCR primers to be used to clone the gene from land plant species. For phylogenetic analysis, the alignment was truncated to include domains F-H of RPB1 (Figure 2), since this was prior to the publication of the sequence for domains A-F from *Dictyostelium* by Stiller and Hall (1998). Figure 3 represents a maximum likelihood phylogenetic tree inferred from amino acid sequences from domain F-H. Neighbor-joining and parsimony trees inferred from the same dataset and from a corresponding alignment of nucleotide sequences yielded very similar trees (data not shown). Analysis of nucleotide composition showed that there was little compositional bias at the first and second codon positions, but that the third positions were biased (data not shown). The consistent and strong statistical support for the relationships in this preliminary analysis indicated that RPB1 would be a good phylogenetic marker with which to study the evolutionary relationships of land plants.

Figure 2: A 606-position amino acid alignment of 13 sequences from RPB1 F-H, with the slime mold *Dictyostelium discoideum* as the outgroup to Animals, Plants and Fungi. Residues identical to that of the outgroup are shown as points (.), while gaps in the alignment are shown as a dash (-). The alignment was prepared with Clustal V with a floating gap penalty of 90 and a fixed gap penalty of 10, with further manual refinement.

```
                                                                                                        100
Dic.dis-II    AMGGREGLIDTAVKTSETGY IQRCLVKAMEDVSIKYDATV RNSLGDVIQFAYGEDGIDGC FVENQSIDSLRKDNTELERM YRHQVDKPDYGDGWMDPLVI
Cae.ele-II    ...............A.... ...R.I....S:MVN..G.. ....AQMV.LR.....L..M W....NMPTMKPN.AVF..D F.VS.AQNAIKLMDLTDNKF
Dro.mel-II    ...............A.... ...R.I....S.MVN..G.. ...V.QL..LR.....LC.E L..F.NMPTVKLS.KSF.KR FK----------FDWSNERL
Art.sal-II    ...............A.... ...R.I....ACMVA..G.. ...V.Q...LR.....LA.E L..F..LPTIKLS.RAF.SK ..----------FDPSNERQ
Cra.gig-II    ...............A.... ...R.I....S.MV...G.. ..QVEQLV.LR.....L.AT H..F.TMPT.KPS.RAF..Q FK----------FDATNERN
Hel.sta-II    ...............A.... ...R.I....S.MV...G.. ..QIEQL..LR.....LA.E W..F.NLP..KPS.KAF.AG FK----------FDPTNEKH
Ily.obs-II    ...............A.... ...R.I....S.MV...G.. ..QVEQLV.LG.....L.A. H..F..LPT.KPSTRAF..Q F.----------FDPTDERM
Mus.mus-II    ...............A.... ...R.I.S..S.MV...... ...INQ.V.LR.....LA.E S..F.NLAT.KPS.KAF.KK F.----------FDYTNERA
Hom.sap-II    ...............A.... ...R.I.S..S.MV...... ...INQ.V.LR.....LA.E S..F.NLAT.KPS.KAF.KK F.----------FDYTNERA
Sac.cer-II    ...............A.... ...R....L..IMVH..N.T .....N....I.....M.AA HI.K..L.TIGGSDAAF.KR ..----------VDLLNTDH
Sch.pom-II    ..A............A.... ...R........MVR..G.. ..AM..I.........L.AT L..Y.VF....LSTKQF.KK ..----------IDLMEDRS
Arab.th-II    .................... ...R.......IMV...G.. ..........L.....M.AV WI.S.KL...KMKKS.FD.T FKYEI.DENWNPTYLSDEHL
Gly.maxC-II   .................... ...R.......IML...G.. ..........L.....M.AI WI.T.KL.T.KMKK..FD.V F.YEF.EENWKPNY.LQEPV
              _________________________________________________________________
                     domain F

                                                                                                        200
Dic.dis-II    EHVRNDSLTRDTLEKEFERI KSDRSLLRNEIIPSGEANWP LPVNLRRLINNAQKLFNIDI RRVSDLNPAVVVLEIEKLVA RLKIIATADTTEDDENFNRA
Cae.ele-II    LRKNYSEDVVREIQESEDG. SLVE.EWSQLEEDRRLLRKD F.RGDAKIVLPCNL.RL.WN AQKIFKVDLRNAVNLSP.HV ISGVRELSKKLIIVSGNDEI
Dro.mel-II    MKKVFTDDVIKEMTDSS.A. QELEAEWDRLVSDRDSLRQI F.NGESKVVLPCNLQRM.WN VQKIFHINKRLPTDLSPIRV IKGVKTLLERCVIVTGND.I
Art.sal-II    LRKTFTEDVTKS.LGDS.V. TEIEKEWEALVKDREALRKV F.SGENKVVLPCNLQRM.WN AQKTFHINKRMPSDLSP.RV IQGVRDLLKNCVIVNGEDKL
Cra.gig-II    MKKCLSEEVIKD.MGDALAV SQLDREWEQLREDRDILRSI F.TGDSKVVLPCNLQRM.WN AQKIFRIDTHKPTDLHPIKV VEGVEDLCKRLIVVAGGD.L
Hel.sta-II    LKNYL.EDILXS.XXDANV. AEVE.EYKQLLEDRTAIRQI F.SGDSKIVLPCNLQRL.WN AQKIFRIHTRKPSNLHPVKI IEDVRELSKKFMIVKGED.L
Ily.obs-II    MKRCLKEDVIKG.RG.HRL. EELE.EWLQL.SDRDS.RQV FSTGDA.IVLPCNLQRM.WN AQKIFRIDKRKPSDLDPINV VSGVRDLCQRLTIV.GEDMI
Mus.mus-II    LRRTLQEDLVKDVLSNAHIQ NELEREFERMREDREVLRVI F.TGDSKVVLPCNL.RM.WN AQKIFHINPRLPSDLHPIKV VEGVKELSKKLVIVNGDDPL
Hom.sap-II    LRRTLQEDLVKDVLSNAHIQ NELEREFERMREDREVLRVI F.TGDSKVVLPCNL.RM.WN AQKIFHINPRLPSDLHPIKV VEGVKELSKKLVIVNGDDPL
Sac.cer-II    TLDPSLLESGSEILGDLKLQ VLLDEEYKQLVKDRKFLREV FVDGEANWPLPVNIRRI.QN AQQTFHIDHTKPSDLTIKDI V.GVKDLQENLLVLRGK.EI
Sch.pom-II    L---SLYMENSIENDSSVQD LL.EEYTQLVADRELLCKFI F.KGDA.WPLPVNVQRI.QN ALQIFHLE.KKPTDLLPSDI INGLNELIAKLTIFRGSD.I
Arab.th-II    .DLKGIREL..VFDA.YSKL ET..FQ.GT..ATN.DST.. ....IK.H.W....T.K..L .KI..MH.VEI.DAVD..QE ..LVVPGD.ALSVEAQK.--
Gly.maxC-II   .DLKTIREF.NVF.A.VQKL EA..HQ.AI..ASN.DNSL. .....K...W....T.KV.F ..P..MH.MEI.EA.D..QE ...VVPGE.ALSQEAQK.--
```

```
                                                                                                   300
Dic.dis-II    WAEVYFNATMLFSILVRSTF ASKRVLTEFRLTEKAFLWVC GEIESKFLQALAHPGEMVGA LAAQSIGEPATQMTLNTFHY AGVSSKNVTLGVPRLNQIIN
Cae.ele-II    SKQAQY...L.MN..L...L CT.NMC.KSK.NSE..D.LL .....R.Q..I.Q....... ....L............... ....A..........KE...
Dro.mel-II    SKQANE...L..QC.I...L CT.Y.SE....STE..E.LV ....TR.Q..Q.N....... ....L..............F ...............KE...
Art.sal-II    SKQANE...L..QC.....L CT..LAD.Y..NTE..E.LI ....TR.QL.Q.V....... I...L..............F ...............KE...
Cra.gig-II    SIQANA...L..KC.I...L CA...TE....SSE..E.LI ..V..R.Q..Q......... ....L............... ....A..........KE...
Hel.sta-II    SKTANT...L.MN......L C....IE..H.STE..E.LM .......I..RVQ....... ....L..............X ....A..........KE...
Ily.obs-II    SKQANE...L.MKC.I...L CA...AE.H..S.D..N.LL ....TR.Q..Q......... ....L............... ....A..........KE...
Mus.mus-II    SRQAQE...L..N.HL...L C.R.MAE....SGE..D.LL .......N..I......... ....L............... ....A..........KEL..
Hom.sap-II    SRQAQE...L..N.HL...L C.R.MAE....SGE..D.LL .......N..I......... ....L............... ....A..........KEL..
Sac.cer-II    IQNAQRD.VT..CC.L..RL .TR...Q.Y...KQ..D..L SN..AQ..RSVV.......V ...................F ...A..K..S.....KE.L.
Sch.pom-II     TRD.QN...L..Q..L..K. .V...IM.Y..NKV..E.IM ..V.AR.Q..VVS......T .................... ...............KE.L.
Arab.th-II    -------..LF.N..L...L ......E.YK.SRER.E..I .....R...S.VA....I.C V................... ....A..........RE...
Gly.maxC-II   -------..L..N..L.... ......E.Y..SRES.E..V .....R...S.VV....I.V V..................F ....A..........REL..
                                                  ____________________________________________________________________
                                                                         domain G

                                                                                                   400
Dic.dis-II    IAKQVKTPSLTIYLKPHMAR DMDRAKIVKSQLEYTTLANV TSATEIYYDPDPQNTIISED AEFVNSYFELPDEEIDVH-- SMSPWLLRIELDRG-MVTDK
Cae.ele-II    VS.TL......VF.TGAA.K .PEK..D.LCK..H..---. .CN.A.......K..V.A.. E.W.SIFY.M.--DH.LS-- RT...........KR..-..
Dro.mel-II    .S.KP.A....VF.TGGA.. .AEK..N.LCR..H...RK. .AN.A.........R.V.... Q....V.Y.M.--DF.PT-- RI...........KR.T-..
Art.sal-II    VSRSP.A....VF.TGTA.. .AEK..Q.LCR..H...KK. .AN.A.....XIK..V.A.. QD..TL.YDM.--DV.TS-- R....X.......KR.T-..
Cra.gig-II    .S.KP......V..IGQA.. .AEK..D.LCR..H...RK. .AN.A..........V.... Q.W..V.Y.M.--DF..S-- KI.A.........KR.T-..
Hel.sta-II    VS.KPRA....VI.IGQP.. .AEK..D.LC...H...RK. .EN.A.......MH.L.E.. Q.W.YI.YDM.--DV.IS-- RL......V....KR.T-..
Ily.obs-II    .S.KP......V..NGPA.. .AEKC.D.LCR..H...RK. .AN.A........M..V.V.. Q.W.SI.Y.M.--DF.AS-- RI...........KR.T-..
Mus.mus-II    .S.KP......VF.LGQS.. .AE...DILCR..H...RK. .AN.A.....N..S.VVA.. Q.W..V.Y.M.--DF..A-- RI......V....KH.T-.R
Hom.sap-II    .S.KP......VF.LGQS.. .AE...DILCR..H...RK. .AN.A.....N..S.VVA.. Q.W..V.Y.M.--DF..A-- RI......V....KH.T-.R
Sac.cer-II    V..NM......V..E.GH.A .QEQ..LIR.AI.H...KS. .I.S........RS.V.P.. E.IIQLH.S.L...AEQSFD QQ......L....AA.N-..
Sch.pom-II    V..NI.........M.WI.A N..L..N.QT.I.H...ST. ......H......D.V.E.. KD..EAF.AI....VEENLY KQ......L....AK.L-..
Arab.th-II    V..RI.....SV..T.EASK SKEG..T.QCA......RS. .Q...VW.....MS...E.. F...R..Y.M...DVSPD-- KI.........N.EM..-..
Gly.maxC-II   V..SI.....SV....DAGK TKE...T.QCA......RS. .Q...VW.....MS...E.. VD..M..Y.M....VALE-- KI...FV.V..N.EM..-..
             _____________________________
                      domain G
```

```
                                                                                                                    500
Dic.dis-II     KLTMADITQCVVRDFGLSLN CIFSDDNAEKLILRIRMVES QETKGT---DND----DDDQ FLRRIESNMLSEMVLRGIKG IKKVFMRTDEKIPK----VT
Cae.ele-II     ....EM.ADRIHGG..NDVH T.YT......VF.L.IAGE DKG---EAQEEQVDKME..V ...C..A....DLT.Q..PA .S..Y.NQPNTDD.KRIII.
Dro.mel-II     ....EQ.AEKINVG..ED.. ...N....D..V....IMNN E.N.F-QDE.EAVDKME..M ...C..A....D.T.Q..EA .G..Y.HLPQTDS.KRIVI.
Art.sal-II     ....EQ.SEKITAG..ED.. ...N.......V....IMN. D.S.F-GEE.EQIDKME..M ...C..A....D.T.Q..EA .S..Y.HLPQTDN.KRIII.
Cra.gig-II     ....EQ.SEKITAG..DD.. ...N.......V....IMN. D.S.M-QNEEEIVDKME..V ...C..A.L..D.T.Q..EA .A..Y.HLPNTED.KRISI.
Hel.sta-II     ....EQ.SEKITAG..DD.. ...N.......V..V.LMSN .DG.QDQDTEEQIDKMP..T ..KH......TD.T.Q..TS .A..Y.QQPTTDD.KRIIID
Ily.obs-II     ....EQ.SEKITAG..DD.. ...N.......V....IMN. DDS.M-QDEEEVVDKME..V ...C....L..D.T.Q..EA .A..Y.HLPTTDD.KRIHI.
Mus.mus-II     ....EQ.AEKINAG..DD.. ...N.......V....IMN. D.N.M-QEEEEVVDKM...V ...C......TD.T.Q..EQ .S..Y.HLPQTDN.KKIII.
Hom.sap-II     ....EQ.AEKINAG..DD.. ...N.......V....IMN. D.N.M-QEEEEVVDKM...V ...C......TD.T.Q..EQ .S..Y.HLPQTDN.KKIII.
Sac.cer-II     D...GQVGERIKQT.KND.F V.W.E..D....I.C.V.R- --P.SLDAETEA----EE.H M.KK..NT..ENIT...VEN .ER.V.MKY-----DRKVPS
Sch.pom-II     ..S.S.VAGKIAES.ERD.F T.W.E...D...I.C.IIRD DDR.AEDDDNMI----EE.V ..KT..GH..ESIS...VPN .TR.Y.MEH-----KIVRQI
Arab.th-II     ..S....AEKINLE.DDD.T ...N....Q......IMND EGP..ELQDESA----E..V ..KK......T..A....PD .N...IK-QVRKSR----FD
Gly.maxC-II    ..S....A.KINLE.DDD.S ...N...........IMND DAP..EVQDESA----E..V ..KK..T...T..T....PD .N...IK-NA.VQ.----FD

                                                                                                                    600
Dic.dis-II     ENSGFGVREEWILDTDGVSL LEVMSHPDVDHTRTTSNDIV EIIQVLGIEAVRNALLKELR AVISFDGSYVNYRHLAILAD VMTYRGHLMAITRHGINRVE
Cae.ele-II     PEG..KSVAD...E...TA. .R.L.ERQI.PV.......C ..FE........K.IER.MD N..............L.C. ...AK.....YS......Q.
Dro.mel-II     .TGE.KAIG..L.E...T.M MK.L.ER...PI..S....C ..F.........KSVE..MN ..LQ.Y.L........L.C. ...AK.............QD
Art.sal-II     DTGE.RAIA..L.E...T.. MK.L.ER...PV..Y....C ..FT........KSIE..MN ..LQ.Y.L........L.C. ...AK.Y...........QD
Cra.gig-II     .EGE.KAVA....E...TA. MK.L.QR...PI...T.... .VFSI.......KSIE..MN H.............SL.C. S..AK.............Q.
Hel.sta-II     .KGE.KALQD...E...TA. RR.L.VEN..PVK.V..... .VFE........KSIER.MN N..............L.C. ...AK.............Q.
Ily.obs-II     .EGE.KAVA....E...S.. MK.L.ER...PV..YT.... .VFAT.......K.IER.MV H..............L.C. ...AK.............Q.
Mus.mus-II     .DGE.KALQ....E...... MR.L.EK...PV........ ..FT........K..ER..Y H..............L.C. T..C...........V..QD
Hom.sap-II     .DGE.KALQ....E...... MR.L.EK...PV........ ..FT........K..ER..Y H..............L.C. T..C...........V..QD
Sac.cer-II     PTGEYVKEP..V.E....N. S...TV.GI.P..IYT.SFI D.ME......G.A..Y..VY N..AS.........M.L.V. ...TQ.G.TSV....F..SN
Sch.pom-II     .DGT.ERAD..V.E...IN. T.A.TVEG..A...Y..SF. ..L.I.....T.S....... N..E............L.C. ...S..............A.
Arab.th-II     .EG..KTS...M...E..N. .A..C.E...PK.....HLI ...E........R...D... V................C. T.................ND
Gly.maxC-II    ..E..KSNG..M...E..N. .A..C.E...A......HLI .V.E........RS..D... I...............CE T.................ND
                                                                                  domain H
```


606
Dic.dis-II TGPLMR
Cae.ele-II V.A...
Dro.mel-II ..A...
Art.sal-II ..A...
Cra.gig-II ..A.A.
Hel.sta-II ..V.A.
Ily.obs-II ..A.A.
Mus.mus-II .....K
Hom.sap-II .....K
Sac.cer-II ..A...
Sch.pom-II ..A...
Arab.th-II ......
Gly.maxC-II ...M..
domain H

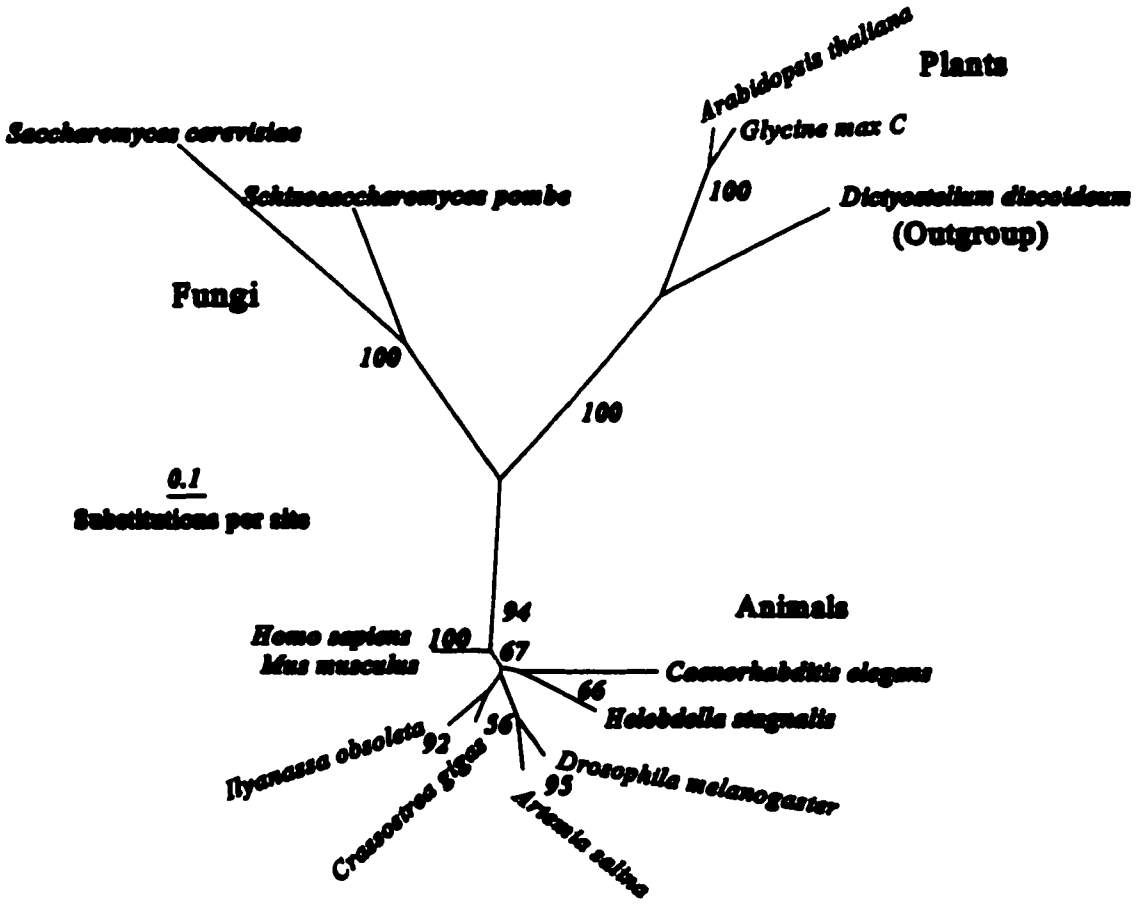


Figure 3: A maximum-likelihood distance phylogenetic tree with non-clocklike branch lengths inferred from a 606-position alignment of RPB1 amino acid sequences of domains F-H from Animals, Plants and Fungi using the JTT model of substitution and an invariable and eight γ-distributed rates of substitution with PUZZLE v4.0.2. The fraction of invariable sites is 0.04 (S.E. 0.03) and the γ parameter α = 0.67 (S.E. 0.06). The log L of this tree is -8954.57, while the unlikely tree with clocklike branch lengths had log L = -8991.17. Numbers at the nodes represent the percent of 10000 quartet puzzling steps in support of that relationship, if >50%.

Plant RPB1:

Using the primers for RPB1 regions D and F which were described in the "Materials and Methods" section, this region was successfully amplified as a 1.3 kb fragment by PCR from genomic DNA of all of the plants listed in table 1 (Figure 4, gel photo). RPB1 D-F PCR products were obtained from both cDNA and genomic DNA templates from *Zea mays*, in which case the nucleotide sequence of the exons was identical. This was confirmed by Southern hybridization (figure not shown) of the PCR products with the *Arabidopsis thaliana* RPB1 clone, pOSA2, and by sequencing clones of nine of these PCR products, as summarized in Table 4. Two clones from *Ephedra viridis*, four clones from maize cDNA and from *Cycas* and *Nymphaea*, and three clones from each of the other plants listed in Table 4 were obtained and sequenced on both strands.

Other clones with sequence similarity to RPB1 (as determined by BLAST scores) from *Araucaria, Marchantia, Ephedra*, and *Equisetum* were obtained and partially sequenced (results not shown), but they were not included in further analysis. Even with very clear sequence data, these sequences could not be unambiguously aligned with other plant RPB1 D-F coding sequences, due to frame shifts in the inferred amino acid sequence and insertions, deletions, or inversions. Some clones obtained from *Marchantia* and *Ephedra* lacked intron positions conserved amongst other plant and green algal RPB1 sequences and had frame shifts and substitutions to other amino acids or stop codons in positions of the inferred amino acid sequences which were perfectly conserved amongst other plant sequences.

PCR amplification of RPB1 using the primer for region A *vs* the region F primer, described in the methods section, was unsuccessful. This was attributed to the abundance of single primer

Figure 4: RPB1 D-F PCR products from genomic DNA and maize cDNA templates, separated with 0.7% agarose in 1xTBE at 100V for 1 hour. Size markers in the first and last lanes are λ DNA digested with *Bst*EII. The arrow corresponds to the major 1.3 kb fragment encoding regions D-F of RPB1. The PCR products (from left to right) are from DNA from the following sources: *Arabidopsis thaliana, Araucaria heterophylla, Cycas revoluta, Ephedra viridis, Equisetum hyemales, Ginkgo biloba, Gnetum gnemon, Magnolia soulangeana, Marchantia polymorpha, Nymphaea odorata, Psilotum nudum, Welwitschia mirabilis, Zea mays*(total DNA), *Zea mays* (cDNA).

Table 4: Identification of RPB1 PCR products from various plants by hybridization with *Arabidopsis* RPB1 and/or from sequence analysis.

Plant	**Southern hybridization signal**	**Sequence**
Araucaria heterophylla	yes	yes
Cycas revoluta	no	yes
Ephedra viridis	no	yes
Ephedra tweediana	no	no
Equisetum hyemales	yes	yes
Ginkgo biloba	yes	yes
Gnetum gnemon	yes	no
Magnolia soulangeana	yes	yes
Marchantia polymorpha	yes	only from probable pseudogenes
Nymphaea odorata	yes	yes
Osmunda claytonea	no	no, but putative RPC1 D-F
Psilotum nudum	yes	no
Welwitschia mirabilis	yes	no
Zea mays	no	yes (identical sequences from cDNA and genomic DNA)

Figure 5: An 834 bp alignment of the RPB1 D-F coding sequence of plants, with *Spirogyra* as the outgroup. Sequences for introns and PCR primers are not shown. The positions of introns 7, 8 and 9 are indicated by a ▼ between the boldfaced nucleotides which flank the intron.

```
                                                                      50
Spirogyra2     CACGTGTGTC AGACATTCGA GACTCGAGCA GAAACCATGG AGCTCATGAT
Arau.het-II    CACGTACCGC AGTCCTTCGA GACTCGGGGC GAAGTAGCCG AGCTCATGCT
Cyc.rev-II     CATGTACCCC AGTCATTTGA GACAAGAGCA GAAGTCTTGG AACTAATGAT
Eph.vir-II     CATGTTCCGC AGTCCTTTGA GACCAGAGCT GAAATCCTGG AGCTGATGAT
Gin.bil-II     CATGTGCCCC AGTCATTTGA GACAAGGGCA GAAGTGTTGG AGTTAATGAT
Mag.sou-II     CATGTTCCTC AGTCATTTGA AACTAGAGCA GAAGTCTTGG AGCTGATGAT
Nym.odo-II     CATGTACCTC AGTGTTTTGA AACTAGAGCA GAAGTCTTGG AGTTGATGAT
Zea mays-2     CATGTCCCCC AGTCATTTGA GACCAGGGCA GAAGTTCTGG AGTTAATGAT
Arab.tha-II    CATGTACCAC AATCATTCGA GACCAGAGCC GAGGTGTTAG AGCTGATGAT

                                                                     100
               GGTGCCAAAA TGTGTCGTGA GTCCCCAATC CAACAGGCCT GTGATGGGCA
               GGTGCCGAAG TGCATCGTCT CGCCGCAGTC GAATCGGCCG GTTATCGGTA
               GGTTCCAAAG TGCATCGTCT CTCCACAGTC GAATAGACCG GTTATGGGTA
               GGTACCAAAA TGTATTGTGT CCCCTCAGTC CAATAGGCCT GTTATGGGTA
               GGTGCCAAAG TGCATTGTTT CTCCAAAGTC CAACAGGCCT GTTATGGGTA
               GGTACCGAAA TGCATTGTCT CGCCTCAATC AAATCGCCCT GTTATGGGTA
               GGTACCAAAA TGTATTGTTT CACCACAGTC CAATCGGCCG GTCATGGGTA
               GGTGCCAAAA TGCATTGTCT CTCCACAATC AAATAGGCCT GTAATGGGTA
               GGTTCCTAAA TGTATTGTCT CCCCCCAGGC GAATCGTCCT GTGATGGGAA

                                                                     150
               TCGTGCAGGA TACGCTACTT GGCTGCAGGA AAGTCACGAA AAGGGACACG
               TCGTGCAAGA CACGCTGCTC GGGTGCCGGA AGGTGACGAA GAGGGATACA
               TTGTCCAAGA CACTCTTTTG GGTTGCAGAA AGATCACAAA GAGAGACACT
               TTGTCCAGGA TACTCTTTTG GGTTGCAGAA AGATCACCAA AAGAGACACA
               TTGTTCAGGA CACTCTTTTG GGTTGCAGAA AGGTGACGAA AAGGGATACA
               TTGTCCAGGA TACACTCTTA GGATGCCGGA AGATAACTAA AAGAGATACC
               TTGTGCAGGA TACACTTCTT GGATGTAGGA AGATCACAAA GCGTGACACC
               TTGTCCAAGA CACACTGCTT GGGTGTCGCA AAATTACTAA AAGGGACACT
               TTGTGCAGGA TACCCTCTTG GGGTGCCGTA AAATTACAAA GAGAGATACT

                          ▼                                          200
               TTTATCGAGA AAGACGTTTT TATGAACATC CTCATGTGGT GGGAAGATTT
               TTCATCGAAA AGGATGTGTT CATGAACATT TTGATGTGGT GGGATGATTT
               TTTATAGAGA AGGATGTGTT CATGAACATC TTAATGTGGT GGGAAGATTT
               TTTATTGAGA AGGATGTTTT CATGAATATC TTGATGTGGT GGGAAGACTT
               TTTATAGAGA AGGATGTCTT CATGAATATC TTAATGTGGT GGGAGGATTT
               TTCATTGAGA AGGATGTCTT TATGAATATC TTGATGTGGT GGGAGGATTT
               TTCATAGAGA AGGATGTTTT CATGAACATT CTGATGTGGT GGGAGGATTT
               CTAATTGAAA AGGATGTATT TATGAACATC TTGATGTGGT GGCAAGATTT
               TTCATAGAGA AGGATGTATT CATGAACACA CTGATGTGGT GGGAAGACTT
```

250

TGAGGGCAAG ATTCCTTCTC CTACGATTTT GAAGCCTCGT CCCCTTTGGA
TGATGGTAAA ATGCCCACCC CAGCGATCCT GAAGCCGAGG CCCATTTGGA
TGATGGCAAA ATACCATCTC CGACTATTCT AAAGCCTAGA CCTCTTTGGA
TGATGGGAAA GTACCTGCAC CTGCAATTTT GAAGCCAAGG CCAATTTGGA
CGATGGCAAA ATACCATCCC CAACAATTCT AAAGCCCAGG CCACTTTGGA
TGATGGGAAA ATACCTGCTC CAACTATTCT GAAGCCTAGA CCTCTTTGGA
TGATGGAAAG ATACCTGCTC CTGCCATTAT GAAGCCTAGA CCTTTGTGGA
CGATGGAAAG ATTCCTGCAC CTACCATTTT GAAACCTAGG CCTATTTGGA
CGATGGGAAA GTTCCGGCTC CTGCAATCTT GAAGCCTCGT CCTCTTTGGA

300

CAGGAAAGCA AGTTTTCTCG TTGATCATCC CCAGAGCTGT GAATCTTGAG
CGGGGAAGCA GATTTTTAGC CTGATCATCC CCAAGCAGAT TAACATGACC
CTGGCAAGCA AGTATTCAAT CTTATCATTC CAAGGCAGAT AAACCTTATA
CTGGCAAGCA AGTTTTTAAT CTTATCATTC CAAAACAAAT AAATCTCGTA
CTGGCAAACA AGTATTCAAT CTTATCATTC CAAGGCAAAT AAATCTAATA
CAGGAAAGCA AGTGTTCAAT CTGATCATTC CTAAGCAGAT AAACCTCATA
CAGGCAAACA AGTATTTAAC CTTATAATTC CAAAGCAGAT AAATTTAATT
CTGGGAAACA AGTTTTTAAC TTAATTATCC CCAAGCAAAT AAATTTAATT
CTGGCAAACA AGTTTTTAAT CTTATCATAC CAAAACAGAT AAATCTGTTG

350

CGGTACTCTG CATGGCATCC CGATTCTGAG AGGGGAGACT TCTCCCCAGG
CGAACTGCGG CATGGCACAA CGACAGTGAG AGTACGGACG TAACGCCCGG
AGATACTCTG CATGGCATTC AGAGTCTGAA ACAGGATTTA TCACGCCAGG
AGGTACTCTG CCTGGCATAA TGAAAGTGAT AGGGGACACA TGACTCTGGG
AGGTACTCTG CATGGCATTC AGAGTCTGAA ACAGGATTTA TCACGCCAGG
AGAACCTCGG CATGGCACTC GGAAGCGGAA ACAGGATTTA TCACTCCAGG
AGGTACTCAG CATGGCATTC GGAATCAGAA ACTGGATTCA TTACCCCTGG
CGGTTTTCAG CATGGCATTC GGAAGAAGAA AAAGGATTTA TTACTCCCGG
AGGTACTCTG CTTGGCACGC AGATACAGAG ACTGGATTTA TAACTCCGGG

400

AGACACTCAA GTGCGTGTGG AGAAGGGAGA GTTGCTCGCA GGAATTCTCT
TGATACCAGC GTTCGCATCG AGAAGGGTGA GCTCCTCACT GGTACGCTCT
GGATACTTGT GTTCGGATTG AAAAGGGAGA AGTTCTGTCG GGCACACTCT
AGATACTGTT GTCAGAATTG AAAAGGGAGA AGTTATAACT GGTACTCTCT
GGATACCTGT GTCCGGATCG AAAAAGGAGA GGTTCTTTCT GGCACTCTGT
AGATACACAA GTTAGAATAG AGAGAGGCGA GCTGCTTGCT GGCACTCTCT
AGATACTTGT GTCCGAATAG AGAGGGGAGA ACTTCTCTCA GGGACCCTCT
TGATACTATG GTCAGGATAG AGAAGGGAGA GCTTCTGTCT GGCACACTTT
GGATACTCAA GTGCGAATTG AAAGAGGGGA ACTTCTTGCC GGAACTCTTT

450

GCAAAAAATC ATTGGGAACC TCCGGTGGAA GTCTCGTTCA CATCATATG**G**
GTAAGAAGAC GCTGGGGACG TCGGGTGGTA GTCTCATTCA CGTGATATG**G**
GTAAAAAAAC CCTTGGAACG TCTTCTGGAA GTCTTATTCA TGTGATCTG**G**
GCAAGAAAAC ACTTGGTGCA TCTAGTGGTA GTCTTATTCA TGTGATATG**G**
GCAAGAAAAC TCTAGGAACG TCTTCTGGGA GTCTGATTCA TGTGATCTG**G**
GCAAGAAGAC CCTTGGAACA TCTACTGGTA GTCTTATTCA TGTCATCTG**G**
GTAAGAAAAC CATGGGCACA TCATCTGGAA GTCTTATTCA CGTTATCTG**G**
GCAAAAAGAG TCTTGGAACA GGCTCCGGAA GTCTTATCCA TGTCATTTG**G**
GCAAAAAGAC CCTTGGTACA TCTAATGGAA GTCTCGTGCA TGTCATTTG**G**

▼ 500

GAAGAAGTAG GTCCTGATGC TGCGCGGAAG TTTCTTGGAC ACACACAATG
GAGGAAGTCG GACCGGACGC TGCCCGTAAG TTCCTCGGTC ACACGCAGTG
GAGGAGGTTG GTCCAGATGC AGCTCGCAAG TTTTTGGGGC ATACACAGTG
GAGGAGGTTG GTCCAGATGC TGCGCGGAAG TTCTTGGGTC ATACGCAATG
GAGGAGGTCG GGCCAGATGC TGCTCGTAAG TTTTTAGGGC ATACACAGTG
GAAGAGGTTG GTCCAGATGC TGCCCGCAAA TTCTTGGGCC ACACACAGTG
GAAGAGGTGG GTCCAGATGC TGCACGCAAG TTTTTGGGGC ATACTCAGTG
GAAGAGGTTG GTCCAGATGC TGCCCGGAAG TTCTTAGGAC ACACACAGTG
GAAGAGGTTG GTCCTGATGC AGCTAGAAAA TTCCTCGGTC ATACTCAATG

550

GCTCGTGAAT TATTGGCTTC TGCAGCAGGG ATTTAGCATC GGAATCGGTG
GCTTGTGAAC TACTGGTTGC TCCAACACGG GTTTAGCATC GGTATCGGTG
GCTTGTTAAC TATTGGCTAT TGCAACAGGG TTTCAGTATT GGTATAGGAG
GCTTGTTAAT TACTGGTTGC TACAACAGGG TTTCAGTATG GGCATTGGAG
GCTTGTCAAC TACTGGCTGT TACAGCAGGG TTTCAGTATT GGTATAGGAG
GCTTGTTAAC TACTGGCTTT TGCAGAATGG ATTTAGTATT GGAATTGGGG
GCTTGTTAAT TATTGGCTTC TGCAGAATGC TTTTAGCATT GGTATTGGGG
GCTTGTAAAC TACTGGCTTC TTCAAAATGG TTTCAGTATT GGAATTGGGG
GCTTGTCAAT TACTGGCTTC TGCAGAATGG TTTTACCATC GGAATTGGTG

600

ATACCATTGC CGATGCATCC ACCATGGATA CCATCAATGA AACCATCGCA
ATACCATCGC CGATTCGGCT ACCATGGAGA AGATCAATGA GACCATCGCT
ACACCATTGC TGATGCTGCA ACAATGGAAA CAATTAATGA AACAATCTCA
ATACAATTGC TGATGCCACA ACAATGGACA CGATCAATGA AACAATACAA
ACACAATTGC TGATGCTGCA ACTATGGAAA CGATTAACGA AACAATTTCA
ACACAATTGC AGATGCATCA ACTATGGAAA AAATTAATGA AACAATATCG
ACACCATTGC AGATGCTTCT ACCATGGAAA AGATTAATGA GACGATCTCT
ATACTATTGC AGATGCATCC ACCATGGAAA CAATTAATGA TACAATATCT
ACACAATTGC CGATTCATCA ACAATGGAGA AAATTAATGA AACTATTTCC

650

AAGGCTAAGA CTGAAGTGAA AGATCTCATC GAAGCAGCTT GTGAGAAACA
AAGGCAAAGC AGGACGTGAA GGAACTGATC AAGGCGTCCC AAGAAAAGAG
AAAGCAAAGG CTGAAGTCAA TCAACTTATT CAGCTTGCTC ACCAGAAAGC
GATGCTAAAA TTAAAGTGCA AGAACTGATA GAAAAGTACA TGGCACACAA
AAAGCAAAGA ATGAGGTCAA ACAACTTATT AAGGCTGCTC AAGAGAAGGC
AAAGCGAAGA ATGAAGTGAA GGAACTTATT AAGGCTGCCC AAGAAAAGCA
AAAGCAAAAA ATGAGGTGAA AGAGCTTATC AAGGCTGCCC AGGAGAAACA
AAAGCTAAGA ATGCTGTGAA GGAGCTTATT AAAAAAGCTC ATGAGAAGCA
AATGCAAAAA CTGCTGTGAA AGATCTTATC CGGCAGTTCC AGGGAAAGGA

700

GTTGGAAGCT CAACCCGGTC GTACCCTCAT GGAGTCCTTT GAGAATCGTG
CCTCGAACCT CAACCCGGTC GCACGCTGAT GGAATCATTC GAGAACAAAG
ATTAGAGGCA GAACCTGGGC GCACGATGAT GGAATCATTT GAAAACAGAG
GCTAGAACAA GAGCCAGGTC GAACGCTATT AGAATCATTT GAAAATCAAG
ATTGGAGGCA GAACCTGGCC GTACAATGAT GGAGTCTTTT GAAAACAGAG
GCTAGAGGCA GAACCTGGGC GAACTATGAT GGAATCATTT GAGAATAGGG
GTTGGAAGCT GAACCTGGGC GAACCATGAT GGAGTCATTT GAGAATAGAG
GTTGGAAGCT GAGCCAGGAC GCACTATGAT GGAATCATTT GAAAACAGAG
ATTGGACCCT GAGCCTGGCC GAACTATGAG AGATACATTT GAGAACAGGG

▼ 750

TCAATCA**GG**T GTTGAACAAG GCCCGTGACG ATGCAGGTCG AGCTGCTCAA
TGAACCA**GG**T TCTGAACAGG GCTCGTGATG AAGCCGGAAG TAGTGCTCAG
TCAATCA**GG**T GTTGAATAAG GCCCGTGATG ATGCAGGAAG TAGTGCTCAA
TCAACCA**GG**T TTTAAATAAG GCTCGAGATG ATGCAGGTAA CAGTGCACAA
TGAATCA**GG**T GTTGAACAAG GCTCGTGATG ATGCGGGAAG TAGTGCTCAA
TGAACCA**GG**T GCTGAATAAG GCTCGTGATG ATGCTGGGAG TAGTGCACAG
TCAATCA**GG**T GTTGAACAAA GCTCGTGATG ACGCTGGTAG TAGTGCTCAG
TGAACCA**GG**T TCTTAACAAA GCCCGTGATG ATGCTGGTAG CAGTGCTCAG
TTAACCA**GG**T TTTGAATAAA GCTCGTGATG ATGCTGGAAG TAGTGCTCAA

800

TCCAGTCTTT CAGAGAGCAA CAACGTCAAA GCCATGGTGA CTGCGGGATC
AAGAGTCTAT CTGAGAGCAA CAACGTCAAG GCGATGGTGA CGGCCGGTTC
AAAAGCTTAT CAGAGAGTAA TAATTTGAAG GCAATGGTTA CTGCTGGCTC
AGAAGTTTAT CTGAGAGCAA TAATTTGAAA GCCATGGTTA CTGCTGGTTC
CGAAGCTTAT CAGAGAGTAA TAATTTGAAG GCAATGGTTA CAGCTGGGTC
AAGAGCTTAT CAGAAAGTAA CAATCTGAAG GCTATGGTGA CAGCAGGATC
AGGAGCTTAT CTGAAAGTAA TAACCTGAAG GCTATGGTCA CTGCAGGATC
AATAGCTTGT CTGAAAGCAA CAATTTGAAG GCTATGGTCA CTGCAGGTTC
AAGAGTTTAG CAGAAACCAA TAACCTTAAG GCCATGGTGA CAGCAGGATC

834

AAAAGGTTCC TTCATCAACA TCTCTCAAAT GATT
CAAGGGATCG TTCATCAACA TCTCGCAGAT GATT
AAAAGGAAGT TTTATTAATA TATCACAGAT GACT
AAAAGGAAGT TTTATCAATA TATCACAGAT GACT
AAAAGGAAGT TTTATTAACA TATCACAGAT GACA
AAAGGGAAGT TTTATCAACA TTTCACAGAT GACT
TAAAGGAAGT TTTATCAATA TTTCACAGAT GACT
AAAAGGCAGT TTCATTAACA TTTCACAAAT GACT
CAAAGGAAGT TTCATCAATA TTTCTCAAAT GACA

artefacts (results not shown) obtained by PCR with the region A primer, which corresponds to a zinc finger domain and so it is likely to be similar to sequences found in genes encoding other proteins which also have DNA-binding domains.

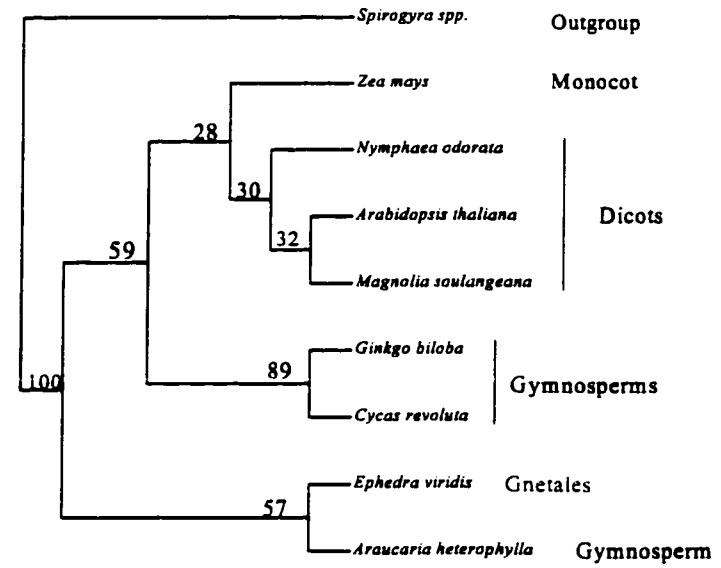
The positions of introns 7, 8 and 9, conserved in *Spirogyra* and *Arabidopsis*, are also conserved in the seven RPB1-coding sequences for regions D-F. This is highlighted in the nucleotide sequence alignment of plant RPB1 D-F coding sequences in Figure 5. In the partial sequence of *Equisetum hyemales* RPB1 D-F, the positions of introns 7 and 9 are conserved as well (results not shown). Dinucleotide and trinucleotide repeats in these introns in *Equisetum hyemales* RPB1 D-F have hampered further sequencing efforts. While intron position is conserved, the intron sequences and lengths vary from species to species amongst plants. Exon sizes, however, are perfectly conserved amongst these plants and *Spirogyra.*

Phylogenetic Analysis of Plant RPB1 D-F DNA Sequences:

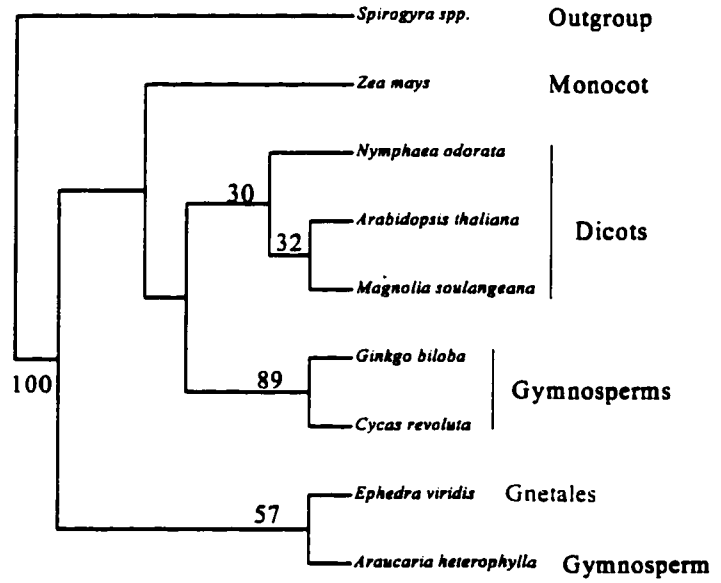
Figures 6, 7 and 8 represent phylogenetic trees inferred from the nucleotide sequences by parsimony, neighbor-joining and maximum-likelihood methods respectively. The alignment used for these analyses was derived from figure 5 and is shown in Appendix 1. The trees in figures 6-8 consistently show *Araucaria*, a Gymnosperm, as one of the most basal taxa. However, two equally parsimonious trees place *Zea mays*, a monocot, as either a sister group to a clade of dicots or to a clade of dicots and gymnosperms (*Gingko* and *Cycas*). In both neighbor-joining and parsimony trees, Gnetales (*Ephedra*) are shown as a sister group to a mixed up clade of angiosperms and the *Cycas/Ginkgo* clade. In figure 7, the neighbor-joining tree shows maize as a sister group to a clade of water lilies (*Nymphaea)* and Gymnosperms (*Cycas* and *Ginkgo).* In figure 8, the PUZZLE Schoeniger-von Haeseler model maximum-likelihood tree resolves fewer relationships that the

previous trees did, with a major difference also being that it grouped Gnetales (*Ephedra*) with an Angiosperm (*Arabidopsis)*, while the parsimony method had placed Gnetales with a Gymnosperm (*Araucaria*) and neighbor-joining placed Gnetales as a sister clade to angiosperms, *Cycas* and *Ginkgo.* albeit with less statistical support. *Cycas* and *Ginkgo* are consistently grouped together by each of the three methods as well. The phylogenetic trees based upon nucleotide sequences, however, are characterized by inconsistencies and low statistical support, if any, for most of the relationships except for the strong support of the *Cycas* and *Ginkgo* clade.

In all of the parsimony and distance phylogenetic trees shown here, no bootstrap values are indicated on the true trees where relationships in the true trees were not seen in the consensus trees obtained after bootstrapping.



A



B

Figure 6: Two equally parsimonious phylogenetic trees of nucleotide sequences encoding plant RPB1 D-F, inferred using DNAPARS in PHYLIP 3.573c. Bootstrap support (% of 1000 replicates) is indicated at the nodes. The third base of codons were removed and the first codon position of the sixfold degenerate arginine and leucine codons were substituted with "m" or "y".

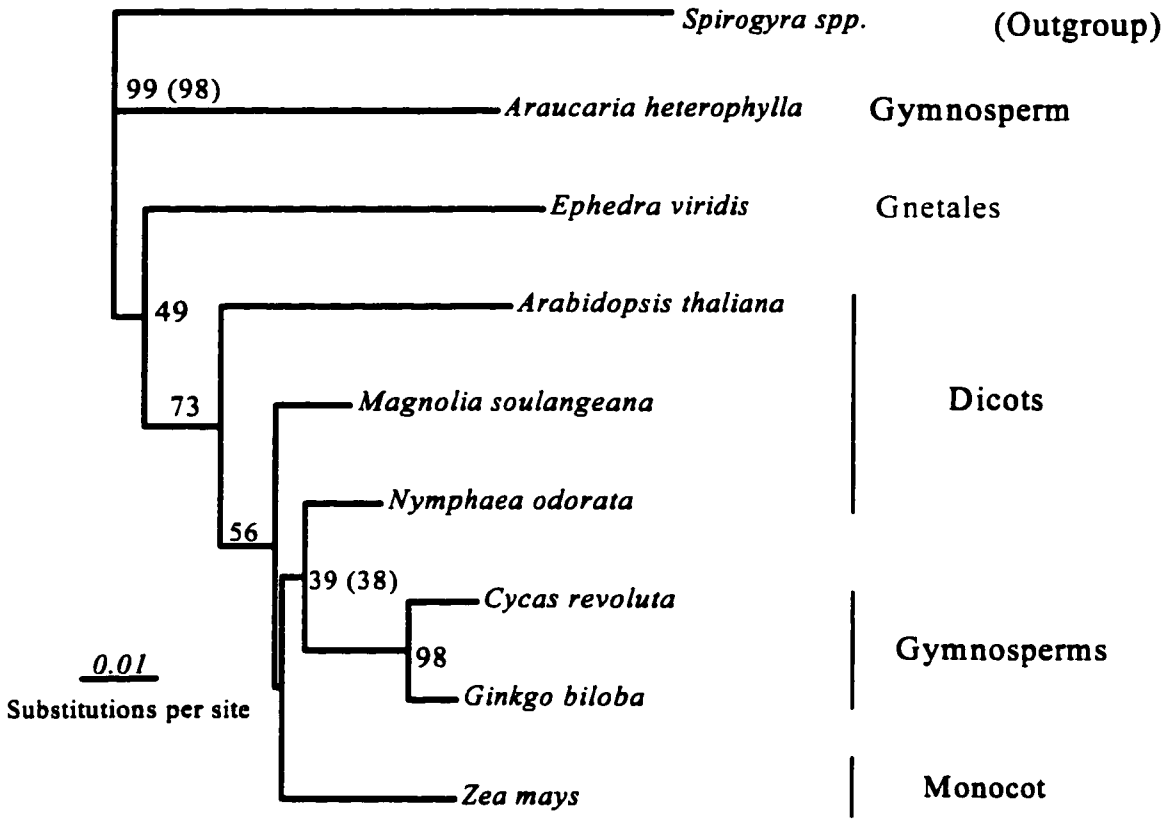


Figure 7: Neighbor-joining phylogenetic tree of nucleotide sequences encoding plant RPB1 D-F. The distances were estimated with DNADIST of PHYLIP 3.573c using the Kimura 2-parameter model of substitution. The % bootstrap values obtained from 1000 replicates are shown at the nodes, with the bootstrap values obtained using the maximum likelihood model of substitution given in parentheses, when different from those obtained by the Kimura model. The third base of codons were removed, and the first base of the sixfold degenerate codons for arginine and leucine were substituted with "m" or "y"

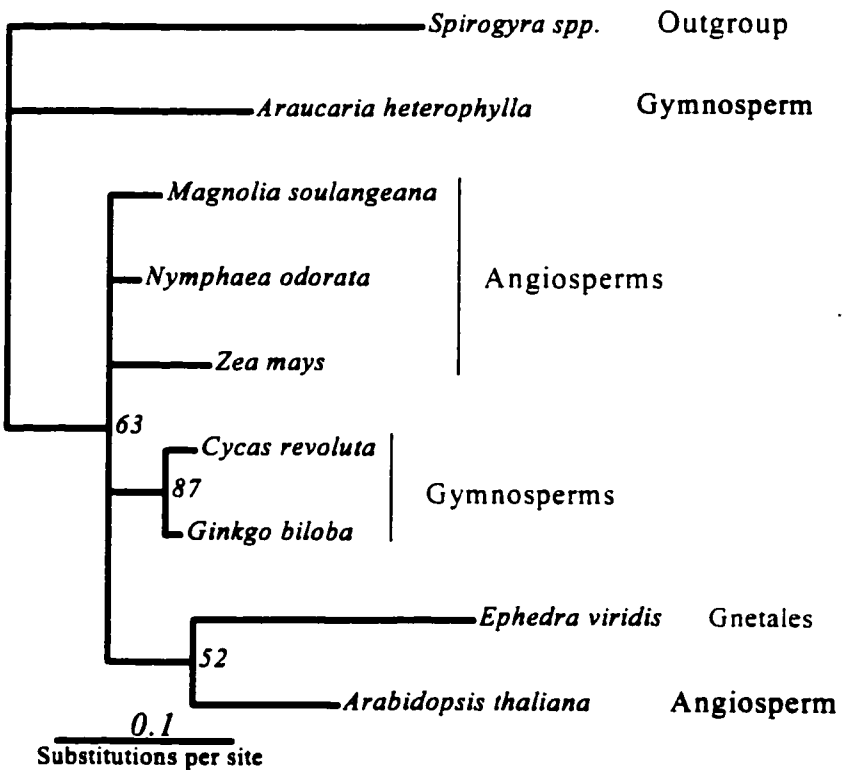


Figure 8: Phylogenetic tree of plant nucleotide sequences encoding RPB1 D-F inferred using maximum likelihood distances with the Schoeniger-von Haeseler model and an invariable and eight γ-distributed rates of substitution, using PUZZLE 4.0.2. The numbers at the nodes indicate the percent of 10000 quartet puzzling steps supporting that relationship, if >50%. The third base of codons were removed and "m" or "y" were substituted for the first base of the sixfold degenerate codons for arginine and leucine. The fraction of invariable sites is 0 (S.E. = 0) and the γ distribution parameter $\alpha = 0.17$ (S.E. 0.03). The log L = -1392.46 for the tree with nonclocklike branchlengths shown here.

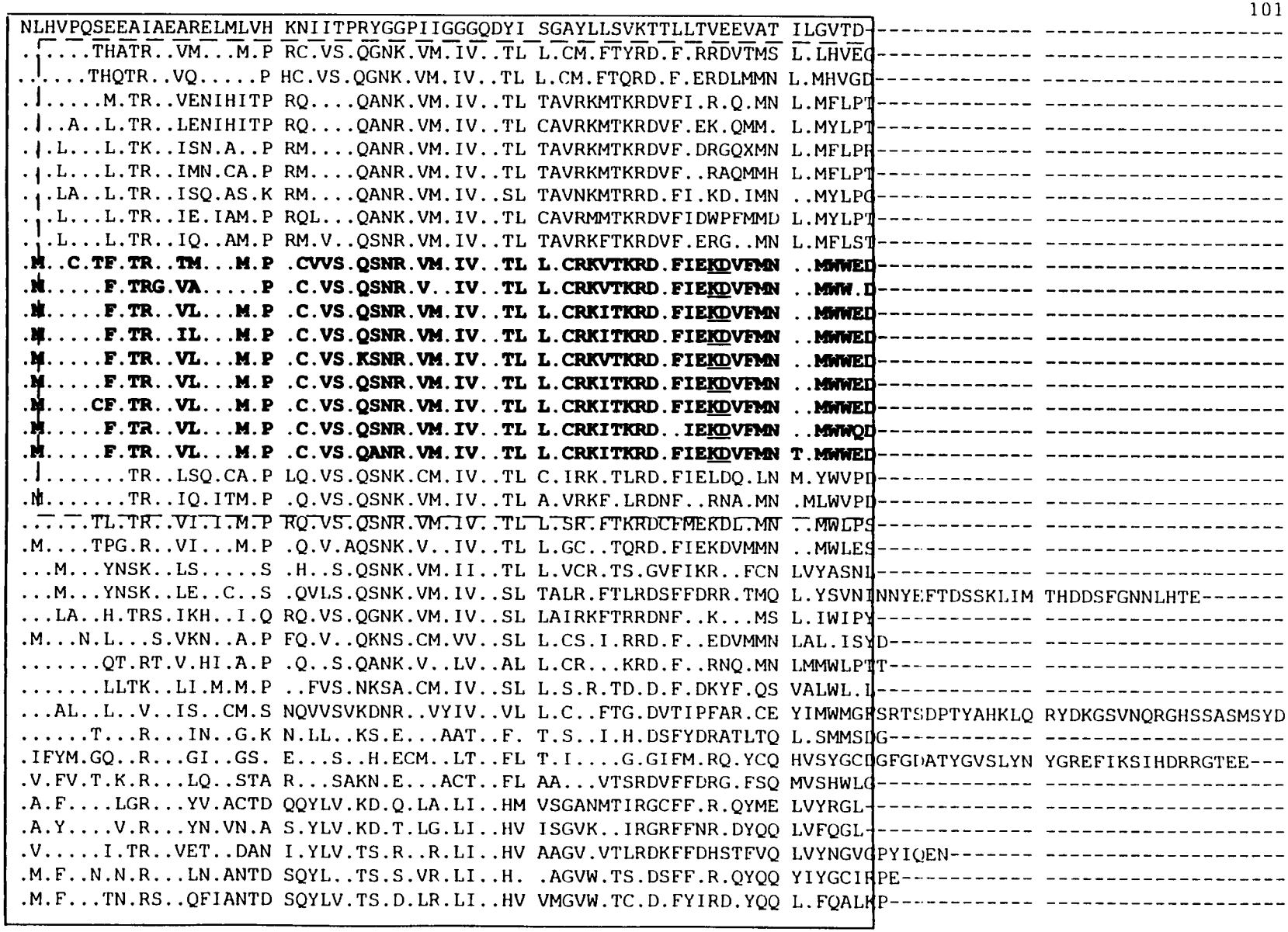
Phylogenetic Analysis of Plant RPB1 D-F Amino Acid Sequences:

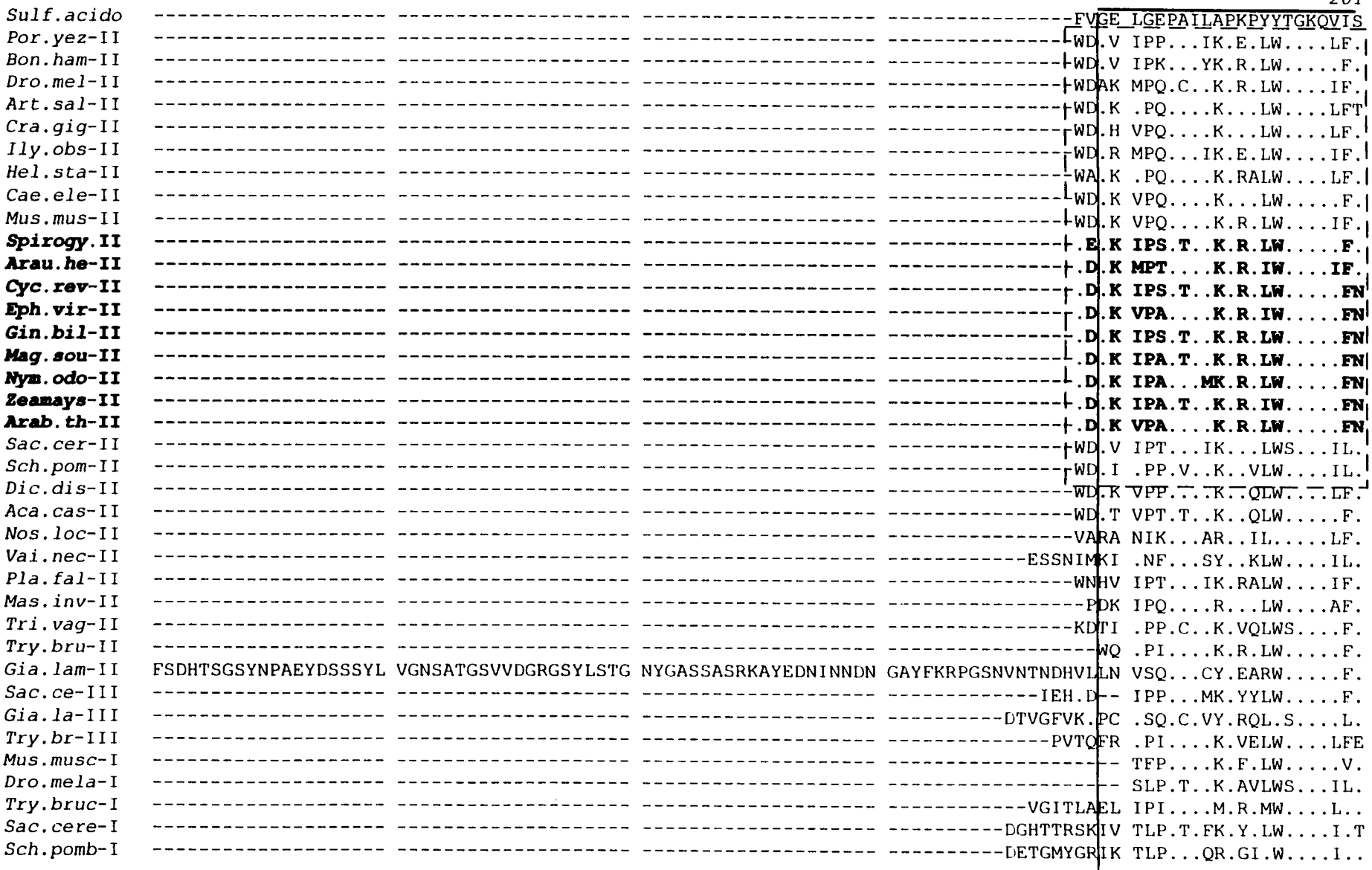
An alignment of the amino acid sequences inferred from the nucleotide sequence alignment in Figure 5 was also used for phylogenetic analysis by the parsimony, neighbor-joining and maximum likelihood methods. The boldfaced section of Figure 9, when considered without any gaps or deletions, represents the amino acid alignment used to infer the phylogenetic trees shown in Figures 10, 11, 12 and 13, with *Spirogyra*, a green alga, as the outgroup (see Appendix 2). Figures 10 and 11 were inferred by the parsimony and neighbor-joining methods, respectively, while Figures 12 and 13 are both maximum likelihood trees, inferred using the Blosum62 and JTT models, respectively.

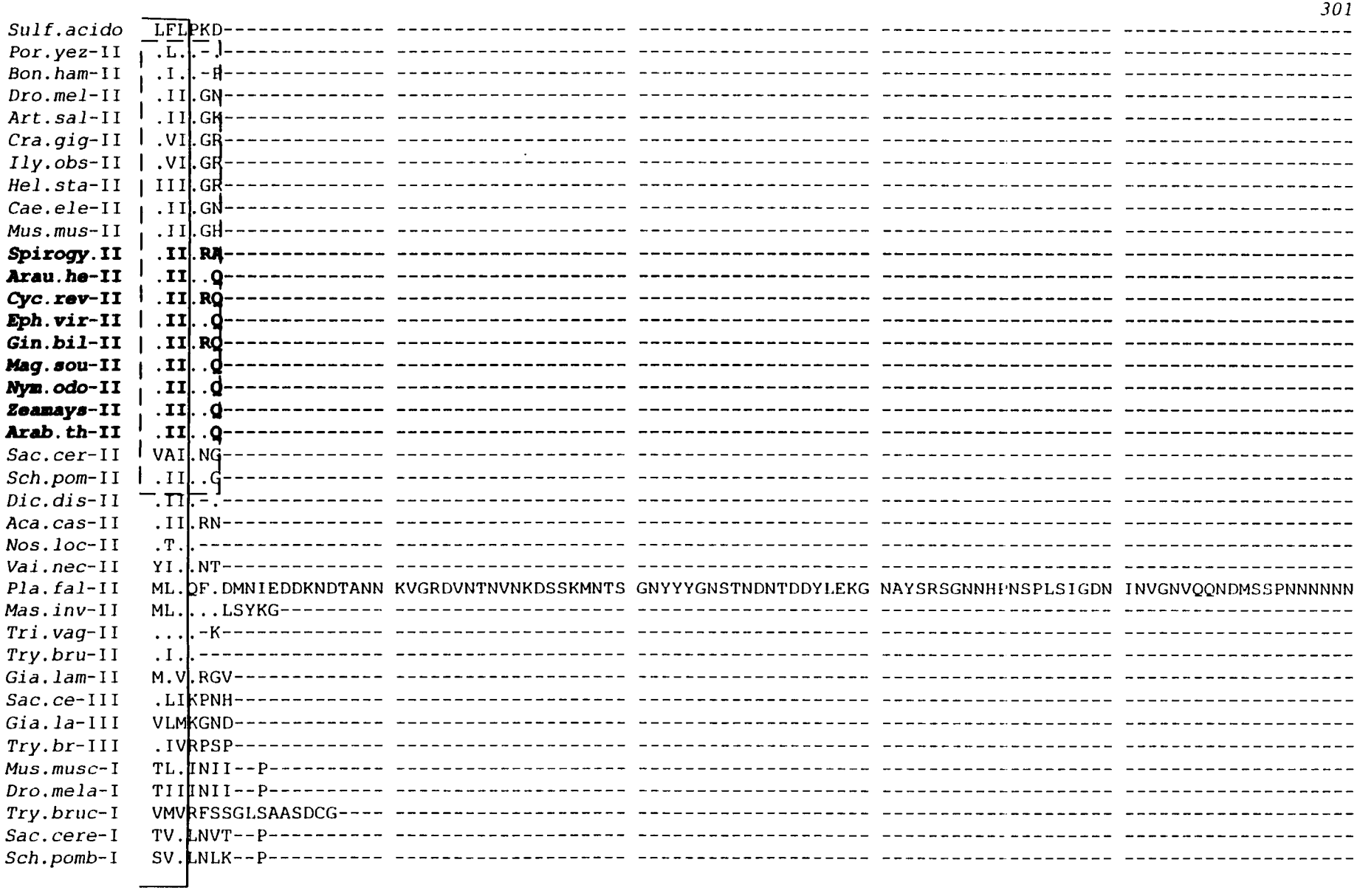
Four equally parsimonious trees are found in Figure 10. These trees consistently group *Cycas* and *Ginkgo* together as sister groups, with strong bootstrap support. Similarly, these trees also consistently place *Araucaria* as the most basal plant and Gnetales (*Ephedra*) as a sister group to angiosperms and the *Cycas/Ginkgo* clade, albeit with low bootstrap support. In three of the four equally parsimonious trees, *Ephedra* is nestled amongst the gymnosperms, and the *Cycas/Ginkgo* clade is a sister group to angiosperms. The parsimony trees are inconsistent about the relationships of angiosperms amongst one another, with monocot, herbaceous or eudicot origins of angiosperms shown as being equally likely.

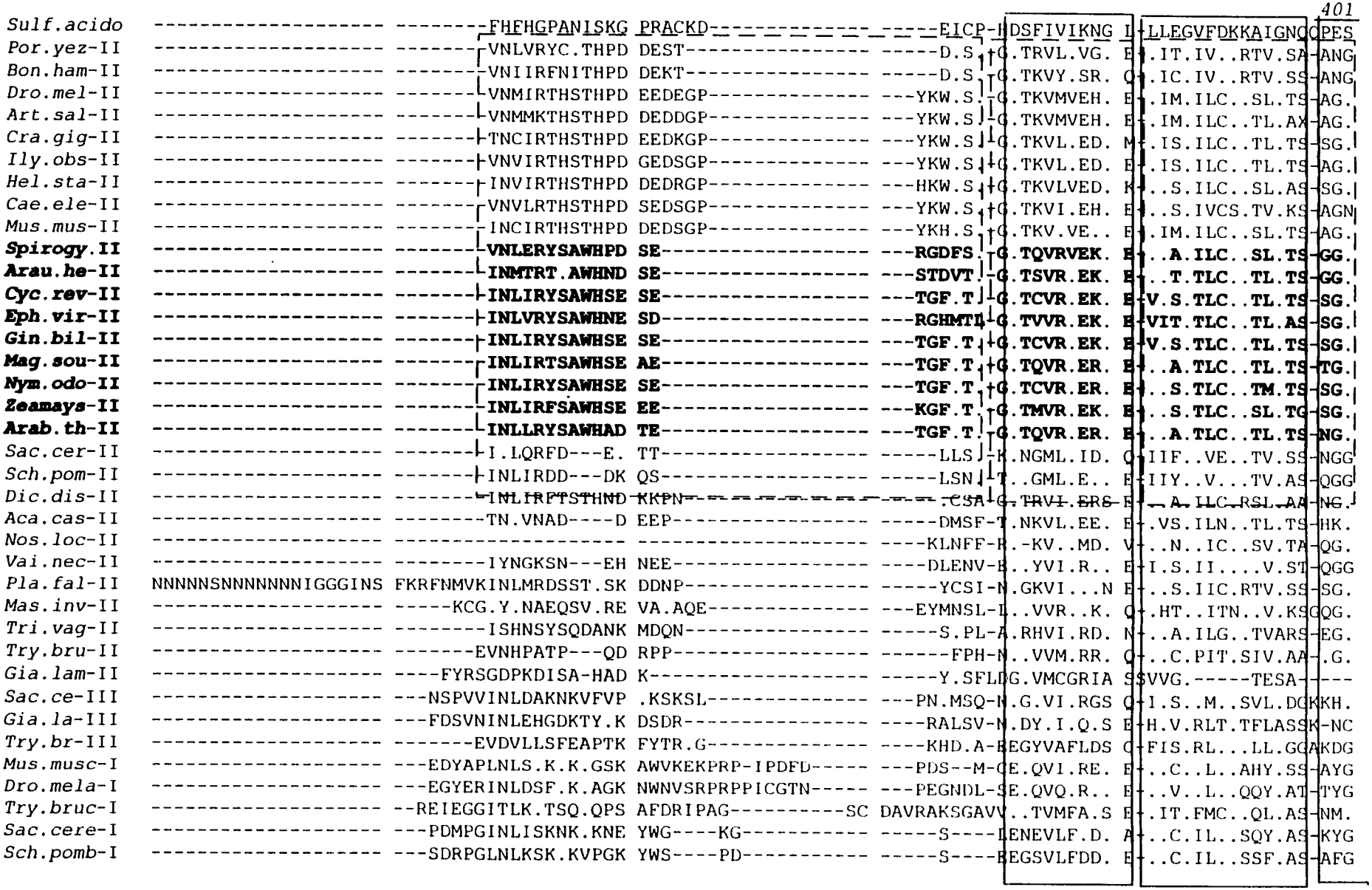
The neighbor-joining tree in Figure 11 shows some relationships similar to those in the parsimony analysis. *Cycas* and *Ginkgo* are grouped together as sister groups, with strong statistical support. *Araucaria* is again placed as the most basal plant, and Gnetales (*Ephedra*) as a sister group to angiosperms and the *Cycas/Ginkgo* clade, with strong statistical support. Relationships amongst angiosperms remain unresolved, except that *Arabidopsis* is placed as basal to both angiosperms and

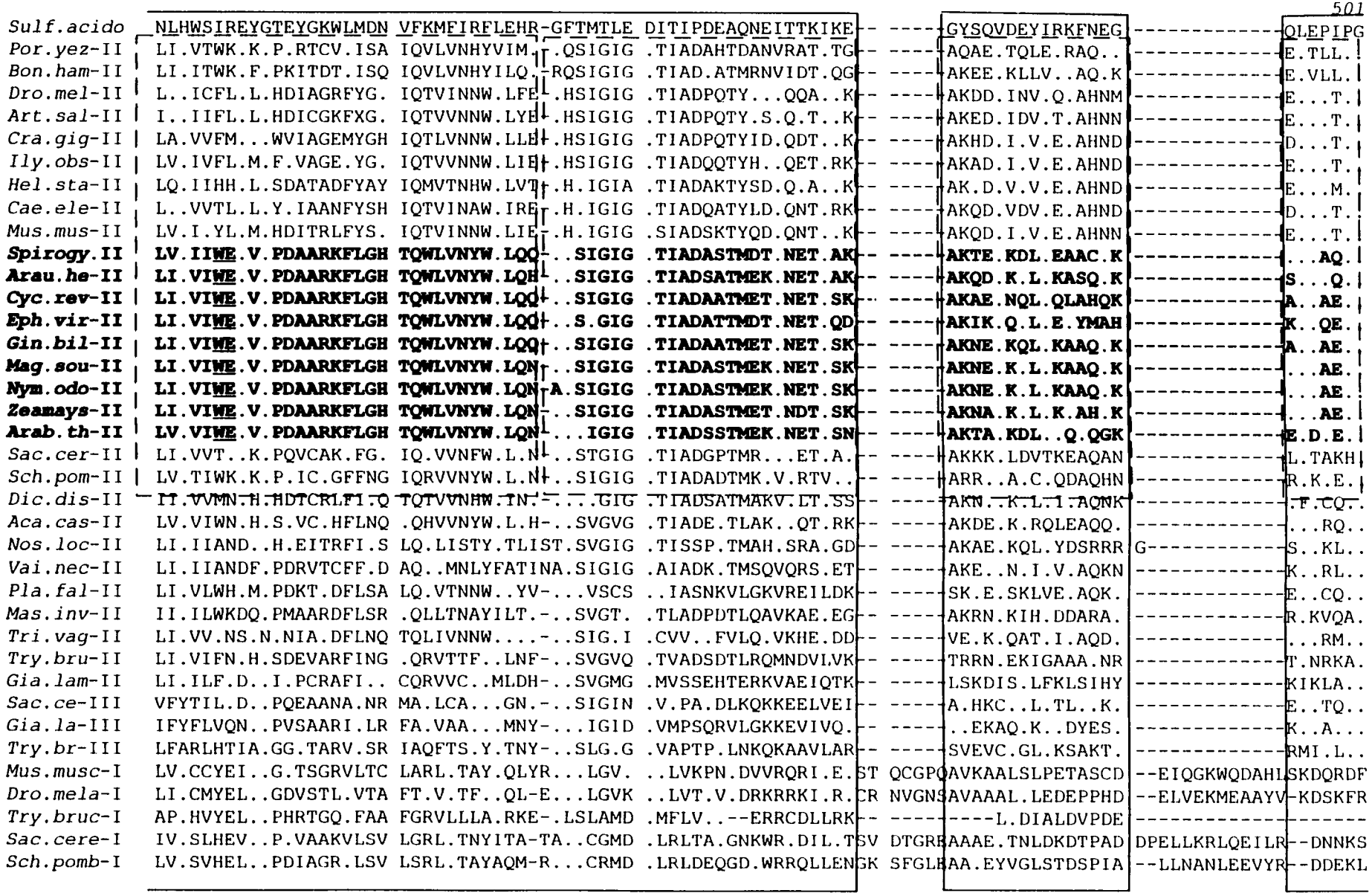
Figure 9: 557 position amino acid alignment of 38 D-F sequences from type A RNA polymerases, with a single Archaeal sequence from *Sulfolobus acidocaldarius* as the outgroup to sequences for eukaryotic RNA polymerases I, II and III. Residues identical to that of the outgroup are shown as points (.), while gaps are shown as a dash (-). Plant sequences are boldfaced. Sequences corresponding to my PCR primers for regions D and F are not shown. Positions of the alignment which were included for phylogenetic analysis of animals, plants and fungi are identified by a box of dotted lines, while positions included for phylogenetic analyses of all of the taxa in the alignment are identified by solid boxes. No gaps or deletions were introduced into the alignment for analyses of plant taxa alone. Residues flanking the three intron positions in plants are underlined. Appendices 2-4 show the alignments used for phylogenetic analysis.

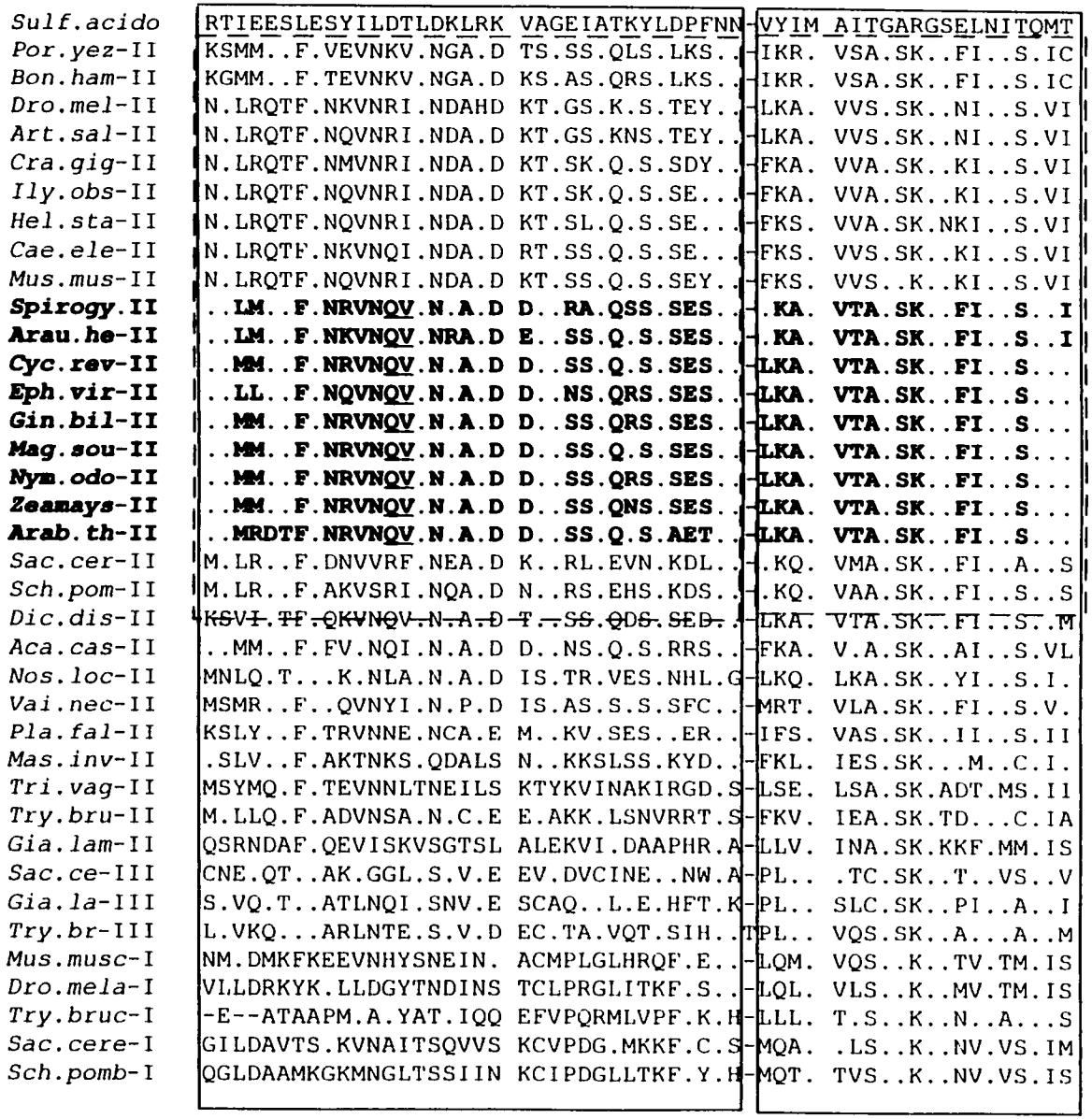



201
Sulf.acido FVGE LGEPAILAPKPYYTGKQVIS
Por.yez-II WD.V IPP...IK.E.LW....LF.
Bon.ham-II WD.V IPK...YK.R.LW.....F.
Dro.mel-II WDAK MPQ.C..K.R.LW....IF.
Art.sal-II WD.K .PQ....K...LW....LFT
Cra.gig-II WD.H VPQ....K...LW....LF.
Ily.obs-II WD.R MPQ...IK.E.LW....IF.
Hel.sta-II WA.K .PQ....K.RALW....LF.
Cae.ele-II WD.K VPQ....K...LW.....F.
Mus.mus-II WD.K VPQ....K.R.LW....IF.
Spirogy.II .E.K IPS.T..K.R.LW.....F.
Arau.he-II .D.K MPT....K.R.IW....IF.
Cyc.rev-II .D.K IPS.T..K.R.LW.....FN
Eph.vir-II .D.K VPA....K.R.IW.....FN
Gin.bil-II .D.K IPS.T..K.R.LW.....FN
Mag.sou-II .D.K IPA.T..K.R.LW.....FN
Nym.odo-II .D.K IPA...MK.R.LW.....FN
Zeamays-II .D.K IPA.T..K.R.IW.....FN
Arab.th-II .D.K VPA....K.R.LW.....FN
Sac.cer-II WD.V IPT...IK...LWS...IL.
Sch.pom-II WD.I .PP.V..K..VLW....IL.
Dic.dis-II WD.K VPP....K..QLW....LF.
Aca.cas-II WD.T VPT.T..K..QLW.....F.
Nos.loc-II VARA NIK...AR..IL.....LF.
Vai.nec-II ESSNIMKI .NF...SY..KLW....IL.
Pla.fal-II WNHV IPT...IK.RALW....IF.
Mas.inv-II PDK IPQ....R...LW....AF.
Tri.vag-II KDTI .PP.C..K.VQLWS....F.
Try.bru-II WQ .PI....K.R.LW.....F.
Gia.lam-II FSDHTSGSYNPAEYDSSSYL VGNSATGSVVDGRGSYLSTG NYGASSASRKAYEDNINNDN GAYFKRPGSNVNTNDHVLLN VSQ...CY.EARW.....F.
Sac.ce-III IEH.D-- IPP...MK.YYLW.....F.
Gia.la-III DTVGFVK.PC .SQ.C.VY.RQL.S....L.
Try.br-III PVTQFR .PI....K.VELW....LFE
Mus.musc-I -- TFP....K.F.LW.....V.
Dro.mela-I -- SLP.T..K.AVLWS...IL.
Try.bruc-I VGITLAEL IPI....M.R.MW....L..
Sac.cere-I DGHTTRSKIV TLP.T.FK.Y.LW....I.T
Sch.pomb-I DETGMYGRIK TLP...QR.GI.W....I..


Sulf.acido LFLPKD----------
Por.yez-II .L..-.----------
Bon.ham-II .I..-P----------
Dro.mel-II .II.GN----------
Art.sal-II .II.GK----------
Cra.gig-II .VI.GR----------
Ily.obs-II .VI.GR----------
Hel.sta-II III.GR----------
Cae.ele-II .II.GN----------
Mus.mus-II .II.GH----------
Spirogy.II .II.RA----------
Arau.he-II .II..Q----------
Cyc.rev-II .II.RQ----------
Eph.vir-II .II..Q----------
Gin.bil-II .II.RQ----------
Mag.sou-II .II..Q----------
Nym.odo-II .II..Q----------
Zeamays-II .II..Q----------
Arab.th-II .II..Q----------
Sac.cer-II VAI.NG----------
Sch.pom-II .II..G----------
Dic.dis-II .TI.-.----------
Aca.cas-II .II.RN----------
Nos.loc-II .T..----------
Vai.nec-II YI..NT----------
Pla.fal-II ML.QF.DMNIEDDKNDTANN KVGRDVNTNVNKDSSKMNTS GNYYYGNSTNDNTDDYLEKG NAYSRSGNNHFNSPLSIGDN INVGNVQQNDMSSPNNNNNN
Mas.inv-II ML....LSYKG---------
Tri.vag-II ....-K----------
Try.bru-II .I..----------
Gia.lam-II M.V.RGV----------
Sac.ce-III .LIKPNH----------
Gia.la-III VLMKGND----------
Try.br-III .IVRPSP----------
Mus.musc-I TL.INII--P----------
Dro.mela-I TIIINII--P----------
Try.bruc-I VMVRFSSGLSAASDCG----
Sac.cere-I TV.LNVT--P----------
Sch.pomb-I SV.LNLK--P----------


401
Sulf.acido
Por.yez-II
Bon.ham-II
Dro.mel-II
Art.sal-II
Cra.gig-II
Ily.obs-II
Hel.sta-II
Cae.ele-II
Mus.mus-II
Spirogy.II
Arau.he-II
Cyc.rev-II
Eph.vir-II
Gin.bil-II
Mag.sou-II
Nym.odo-II
Zeamays-II
Arab.th-II
Sac.cer-II
Sch.pom-II
Dic.dis-II
Aca.cas-II
Nos.loc-II
Vai.nec-II
Pla.fal-II
Mas.inv-II
Tri.vag-II
Try.bru-II
Gia.lam-II
Sac.ce-III
Gia.la-III
Try.br-III
Mus.musc-I
Dro.mela-I
Try.bruc-I
Sac.cere-I
Sch.pomb-I


501
Sulf.acido NLHWSIREYGTEYGKWLMDN VFKMFIRFLEHR-GFTMTLE DITIPDEAQNEITTKIKE---- GYSQVDEYIRKFNEG---------- QLEPIPG
Por.yez-II LI.VTWK.K.P.RTCV.ISA IQVLVNHYVIM.-.QSIGIG .TIADAHTDANVRAT.TG---- AQAE.TQLE.RAQ..---------- E.TLL.
Bon.ham-II LI.ITWK.F.PKITDT.ISQ IQVLVNHYILQ.-RQSIGIG .TIAD.ATMRNVIDT.QG---- AKEE.KLLV..AQ.K---------- E.VLL.
Dro.mel-II L..ICFL.L.HDIAGRFYG. IQTVINNW.LFE-.HSIGIG .TIADPQTY...QQA..K---- AKDD.INV.Q.AHNM---------- E...T.
Art.sal-II I..IIFL.L.HDICGKFXG. IQTVVNNW.LYE-.HSIGIG .TIADPQTY.S.Q.T..K---- AKED.IDV.T.AHNN---------- E...T.
Cra.gig-II LA.VVFM...WVIAGEMYGH IQTLVNNW.LLE-.HSIGIG .TIADPQTYID.QDT..K---- AKHD.I.V.E.AHND---------- D...T.
Ily.obs-II LV.IVFL.M.F.VAGE.YG. IQTVVNNW.LIE-.HSIGIG .TIADQQTYH..QET.RK---- AKAD.I.V.E.AHND---------- E...T.
Hel.sta-II LQ.IIHH.L.SDATADFYAY IQMVTNHW.LVT-.H.IGIA .TIADAKTYSD.Q.A..K---- AK.D.V.V.E.AHND---------- E...M.
Cae.ele-II L..VVTL.L.Y.IAANFYSH IQTVINAW.IRE-.H.IGIG .TIADQATYLD.QNT.RK---- AKQD.VDV.E.AHND---------- D...T.
Mus.mus-II LV.I.YL.M.HDITRLFYS. IQTVINNW.LIE-.H.IGIG .SIADSKTYQD.QNT..K---- AKQD.I.V.E.AHNN---------- E...T.
Spirogy.II LV.IIWE.V.PDAARKFLGH TQWLVNYW.LQQ-..SIGIG .TIADASTMDT.NET.AK---- AKTE.KDL.EAAC.K---------- ...AQ.
Arau.he-II LI.VIWE.V.PDAARKFLGH TQWLVNYW.LQH-..SIGIG .TIADSATMEK.NET.AK---- AKQD.K.L.KASQ.K---------- S...Q.
Cyc.rev-II LI.VIWE.V.PDAARKFLGH TQWLVNYW.LQQ-..SIGIG .TIADAATMET.NET.SK---- AKAE.NQL.QLAHQK---------- A..AE.
Eph.vir-II LI.VIWE.V.PDAARKFLGH TQWLVNYW.LQQ-..S.GIG .TIADATTMDT.NET.QD---- AKIK.Q.L.E.YMAH---------- K..QE.
Gin.bil-II LI.VIWE.V.PDAARKFLGH TQWLVNYW.LQQ-..SIGIG .TIADAATMET.NET.SK---- AKNE.KQL.KAAQ.K---------- A..AE.
Mag.sou-II LI.VIWE.V.PDAARKFLGH TQWLVNYW.LQN-..SIGIG .TIADASTMEK.NET.SK---- AKNE.K.L.KAAQ.K---------- ...AE.
Nym.odo-II LI.VIWE.V.PDAARKFLGH TQWLVNYW.LQN-A.SIGIG .TIADASTMEK.NET.SK---- AKNE.K.L.KAAQ.K---------- ...AE.
Zeamays-II LI.VIWE.V.PDAARKFLGH TQWLVNYW.LQN-..SIGIG .TIADASTMET.NDT.SK---- AKNA.K.L.K.AH.K---------- ...AE.
Arab.th-II LV.VIWE.V.PDAARKFLGH TQWLVNYW.LQN-...IGIG .TIADSSTMEK.NET.SN---- AKTA.KDL..Q.QGK---------- E.D.E.
Sac.cer-II LI.VVT..K.PQVCAK.FG. IQ.VVNFW.L.N-..STGIG .TIADGPTMR...ET.A.---- AKKK.LDVTKEAQAN---------- L.TAKH
Sch.pom-II LV.TIWK.K.P.IC.GFFNG IQRVVNYW.L.N-..SIGIG .TIADADTMK.V.RTV..---- ARR..A.C.QDAQHN---------- R.K.E.
Dic.dis-II II.VVMN.H.HDTCRLFI.Q TQTVVNHW.IN.-...GIG .TIADSATMAKV.LT.SS---- AKN..K.L.I.AQNK---------- .F.CQ.
Aca.cas-II LV.VIWN.H.S.VC.HFLNQ .QHVVNYW.L.H-..SVGVG .TIADE.TLAK..QT.RK---- AKDE.K.RQLEAQQ.---------- ...RQ..
Nos.loc-II LI.IIAND..H.EITRFI.S LQ.LISTY.TLIST.SVGIG .TISSP.TMAH.SRA.GD---- AKAE.KQL.YDSRRR G---------- S..KL..
Vai.nec-II LI.IIANDF.PDRVTCFF.D AQ..MNLYFATINA.SIGIG .AIADK.TMSQVQRS.ET---- AKE..N.I.V.AQKN---------- K..RL..
Pla.fal-II LI.VLWH.M.PDKT.DFLSA LQ.VTNNW..YV-...VSCS ..IASNKVLGKVREILDK---- SK.E.SKLVE.AQK.---------- E..CQ..
Mas.inv-II II.ILWKDQ.PMAARDFLSR .QLLTNAYILT.-..SVGT. .TLADPDTLQAVKAE.EG---- AKRN.KIH.DDARA.---------- R.KVQA.
Tri.vag-II LI.VV.NS.N.NIA.DFLNQ TQLIVNNW....-..SIG.I .CVV..FVLQ.VKHE.DD---- VE.K.QAT.I.AQD.---------- ...RM..
Try.bru-II LI.VIFN.H.SDEVARFING .QRVTTF..LNF-..SVGVQ .TVADSDTLRQMNDVI.VK---- TRRN.EKIGAAA.NR---------- T.NRKA.
Gia.lam-II LI.ILF.D..I.PCRAFI.. CQRVVC..MLDH-..SVGMG .MVSSEHTERKVAEIQTK---- LSKDIS.LFKLSIHY---------- KIKLA..
Sac.ce-III VFYTIL.D..PQEAANA.NR MA.LCA...GN.-..SIGIN .V.PA.DLKQKKEELVEI---- A.HKC..L.TL..K.---------- E..TQ..
Gia.la-III IFYFLVQN..PVSAARI.LR FA.VAA...MNY-...IGID .VMPSQRVLGKKEVIVQ.---- ..EKAQ.K..DYES.---------- K..A...
Try.br-III LFARLHTIA.GG.TARV.SR IAQFTS.Y.TNY-..SLG.G .VAPTP.LNKQKAAVLAR---- SVEVC.GL.KSAKT.---------- RMI.L..
Mus.musc-I LV.CCYEI..G.TSGRVLTC LARL.TAY.QLYR...LGV. ..LVKPN.DVVRQRI.E.ST QCGPQAVKAALSLPETASCD --EIQGKWQDAHLSKDQRDF
Dro.mela-I LI.CMYEL..GDVSTL.VTA FT.V.TF..QL-E...LGVK ..LVT.V.DRKRRKI.R.CR NVGNSAVAAAL.LEDEPPHD --ELVEKMEAAYV-KDSKFR
Try.bruc-I AP.HVYEL..PHRTGQ.FAA FGRVLLLA.RKE-.LSLAMD .MFLV..--ERRCDLLRK---- ----L.DIALDVPDE----------
Sac.cere-I IV.SLHEV..P.VAAKVLSV LGRL.TNYITA-TA..CGMD .LRLTA.GNKWR.DIL.TSV DTGRPAAAE.TNLDKDTPAD DPELLKRLQEILR--DNNKS
Sch.pomb-I LV.SVHEL..PDIAGR.LSV LSRL.TAYAQM-R...CRMD .LRLDEQGD.WRRQLLENGK SFGLPAA.EYVGLSTDSPIA --LLNANLEEVYR--DDEKL


Sulf.acido RTIEESLESYILDTLDKLRK VAGEIATKYLDPFNN-VYIM AITGARGSELNITQMT
Por.yez-II KSMM..F.VEVNKV.NGA.D TS.SS.QLS.LKS..-IKR. VSA.SK..FI..S.IC
Bon.ham-II KGMM..F.TEVNKV.NGA.D KS.AS.QRS.LKS..-IKR. VSA.SK..FI..S.IC
Dro.mel-II N.LRQTF.NKVNRI.NDAHD KT.GS.K.S.TEY..-LKA. VVS.SK..NI..S.VI
Art.sal-II N.LRQTF.NQVNRI.NDA.D KT.GS.KNS.TEY..-LKA. VVS.SK..NI..S.VI
Cra.gig-II N.LRQTF.NMVNRI.NDA.D KT.SK.Q.S.SDY..-FKA. VVA.SK..KI..S.VI
Ily.obs-II N.LRQTF.NQVNRI.NDA.D KT.SK.Q.S.SE...-FKA. VVA.SK..KI..S.VI
Hel.sta-II N.LRQTF.NQVNRI.NDA.D KT.SL.Q.S.SE...-FKS. VVA.SK.NKI..S.VI
Cae.ele-II N.LRQTF.NKVNQI.NDA.D RT.SS.Q.S.SE...-FKS. VVS.SK..KI..S.VI
Mus.mus-II N.LRQTF.NQVNRI.NDA.D KT.SS.Q.S.SEY..-FKS. VVS..K..KI..S.VI
Spirogy.II ..LM..F.NRVNQV.N.A.D D..RA.QSS.SES..-.KA. VTA.SK..FI..S..I
Arau.he-II ..LM..F.NKVNQV.NRA.D E..SS.Q.S.SES..-.KA. VTA.SK..FI..S..I
Cyc.rev-II ..MM..F.NRVNQV.N.A.D D..SS.Q.S.SES..-LKA. VTA.SK..FI..S...
Eph.vir-II ..LL..F.NQVNQV.N.A.D D..NS.QRS.SES..-LKA. VTA.SK..FI..S...
Gin.bil-II ..MM..F.NRVNQV.N.A.D D..SS.QRS.SES..-LKA. VTA.SK..FI..S...
Mag.sou-II ..MM..F.NRVNQV.N.A.D D..SS.Q.S.SES..-LKA. VTA.SK..FI..S...
Nym.odo-II ..MM..F.NRVNQV.N.A.D D..SS.QRS.SES..-LKA. VTA.SK..FI..S...
Zeamays-II ..MM..F.NRVNQV.N.A.D D..SS.QNS.SES..-LKA. VTA.SK..FI..S...
Arab.th-II ..MRDTF.NRVNQV.N.A.D D..SS.Q.S.AET..-LKA. VTA.SK..FI..S...
Sac.cer-II M.LR..F.DNVVRF.NEA.D K..RL.EVN.KDL...-.KQ. VMA.SK..FI..A..S
Sch.pom-II M.LR..F.AKVSRI.NQA.D N..RS.EHS.KDS...-.KQ. VAA.SK..FI..S..S
Dic.dis-II KSVI.TF.QKVNQV.N.A.D T..SS.QDS.SED..-LKA. VTA.SK..FI..S..M
Aca.cas-II ..MM..F.FV.NQI.N.A.D D..NS.Q.S.RRS..-FKA. V.A.SK..AI..S.VL
Nos.loc-II MNLQ.T...K.NLA.N.A.D IS.TR.VES.NHL.G-LKQ. LKA.SK..YI..S.I.
Vai.nec-II MSMR..F..QVNYI.N.P.D IS.AS.S.S.SFC..-MRT. VLA.SK..FI..S.V.
Pla.fal-II KSLY..F.TRVNNE.NCA.E M..KV.SES..ER...-IFS. VAS.SK..II..S.II
Mas.inv-II .SLV..F.AKTNKS.QDALS N..KKSLSS.KYD..-FKL. IES.SK...M..C.I.
Tri.vag-II MSYMQ.F.TEVNNLTNEILS KTYKVINAKIRGD.S-LSE. LSA.SK.ADT.MS.II
Try.bru-II M.LLQ.F.ADVNSA.N.C.E E.AKK.LSNVRRT.S-FKV. IEA.SK.TD...C.IA
Gia.lam-II QSRNDAF.QEVISKVSGTSL ALEKVI.DAAPHR.A-LLV. INA.SK.KKF.MM.IS
Sac.ce-III CNE.QT..AK.GGL.S.V.E EV.DVCINE..NW.A-PL.. .TC.SK..T..VS..V
Gia.la-III S.VQ.T..ATLNQI.SNV.E SCAQ..L.E.HFT.K-PL.. SLC.SK..PI..A..I
Try.br-III L.VKQ...ARLNTE.S.V.D EC.TA.VQT.SIH..TPL.. VQS.SK..A...A..M
Mus.musc-I NM.DMKFKEEVNHYSNEIN. ACMPLGLHRQF.E..-LQM. VQS..K..TV.TM.IS
Dro.mela-I VLLDRKYK.LLDGYTNDINS TCLPRGLITKF.S..-LQL. VLS..K..MV.TM.IS
Try.bruc-I -E--ATAAPM.A.YAT.IQQ EFVPQRMLVPF.K.H-LLL. T.S..K..N..A...S
Sac.cere-I GILDAVTS.KVNAITSQVVS KCVPDG.MKKF.C.S-MQA. .LS..K..NV.VS.IM
Sch.pomb-I QGLDAAMKGKMNGLTSSIIN KCIPDGLLTKF.Y.H-MQT. TVS..K..NV.VS.IS

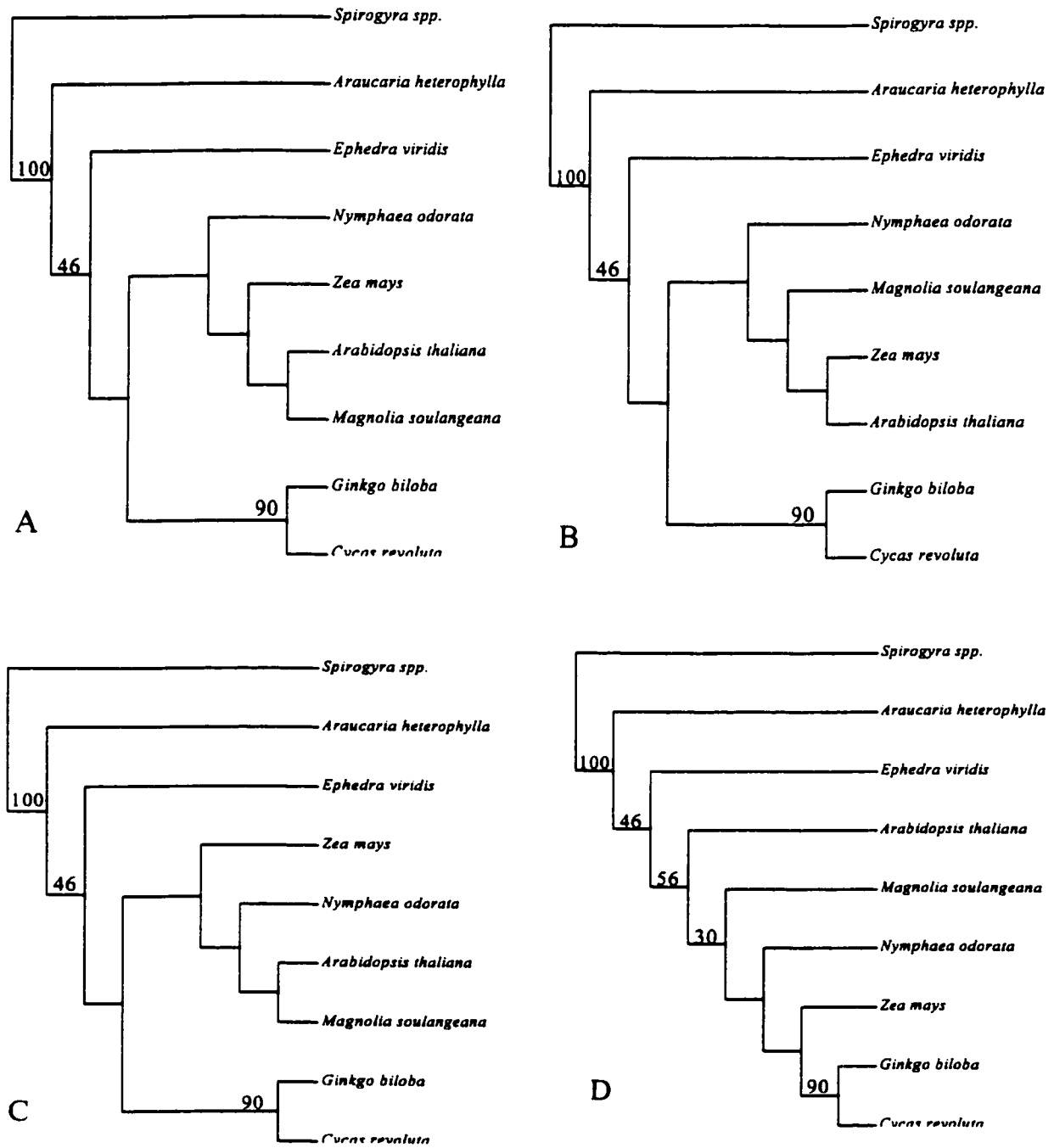


Figure 10: Four equally parsimonious phylogenetic trees inferred from a 280-position alignment of plant RPB1 D-F amino acid sequences using PROTPARS in PHYLIP 3.573c. Bootstrap support (% of 1000 replicates) is indicated at the nodes.

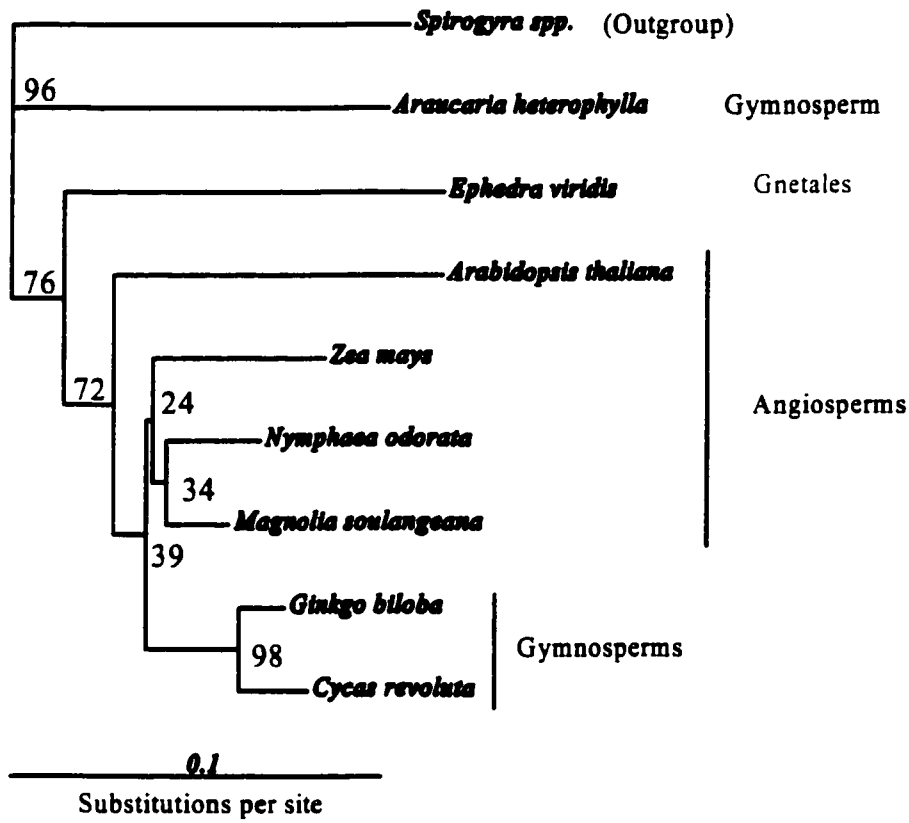


Figure 11: Neighbor-joining distance phylogenetic tree inferred from a 280-position alignment of plant RPB1 D-F amino acid sequences using the Dayhoff PAM model with PROTDIST and NEIGHBOR in PHYLIP 3.573c. Bootstrap support (% of 1000 replicates) is indicated at the nodes.

the *Cycas/Ginkgo* clade, with strong bootstrap support, which is consistent with the neighbor-joining analysis of nucleotide sequences and with only one of the equally parsimonious amino acid trees. In contrast, the two equally parsimonious trees from nucleotide data both placed *Zea mays*, a monocot, as the most basal angiosperm clade. Relationships amongst *Zea mays*, *Nymphaea*, and *Magnolia*, or between these three and their gymnosperm "sister group", are not resolved with bootstrap values greater than 50%. While this neighbor-joining analysis provided greater statistical support for the relationships which it provided, the relationships of most angiosperms are still unresolved.

Figure 12 (A and B) represent trees with non-clocklike and clocklike branch lengths obtained with PUZZLE using the Blosum62 model of maximum-likelihood analysis. In these trees, dicots are all grouped together, with *Arabidopsis* as a sister group to a clade that consists of the woody and herbaceous dicots. *Cycas* and *Ginkgo* are also placed together with strong statistical support, although the exact relationship of this clade to the dicot clade and the monocot is unresolved. Again, Gnetales are placed at the base of the clade consisting of angiosperms and the *Cycas/Ginkgo* clade with strong statistical support and *Araucaria* is placed as the most basal plant lineage. A most basal angiosperm clade is not established from this analysis. The clocklike branch lengths shown in Figure 12B were rejected by PUZZLE, thus considered to be less likely than the non-clocklike branch lengths. In other words, PUZZLE found it unlikely that all of the sequences had identical substitution rates relative to one another. The analysis represented in Figure 12 (A and B) is characterized by its greater statistical support for relationships amongst the ingroups than was found in the previous phylogenetic analyses (Figures 6-8, 10, 11).

Figure 13 (A and B) represent trees with non-clocklike and clocklike branch lengths obtained with PUZZLE using the JTT model of maximum-likelihood analysis. In this set of trees, angiosperms are clearly separated from the remaining taxa with moderate to high statistical support, and the Gnetales (*Ephedra*) are found near the base of the tree, amongst the gymnosperm taxa, with high statistical support. *Zea mays*, a monocot, is found as the most basal angiosperm (i.e., a sister group to dicots), and the *Cycas/Ginkgo* clade is found as a sister group to angiosperms. Also, *Nymphaea*, a herbaceous dicot, is shown as a sister group to *Magnolia* and *Arabidopsis*. The clocklike branch lengths shown in Figure 13B were rejected by PUZZLE in this analysis as well. The relationships shown from this analysis are markedly different from the maximum-likelihood analysis of nucleotide data in Figure 8, which showed Gnetales as a sister group to *Arabidopsis*, an angiosperm, and did not resolve the relationships amongst angiosperms. However, this analysis did consistently place *Araucaria* as the most basal plant, and place *Cycas* and *Ginkgo* as sister taxa to one another. While some parsimony methods did group angiosperms together, Figure 13 represents the only analysis of these sequences in which the angiosperm taxa are grouped exclusively together and also related to one another with moderate to high statistical support.

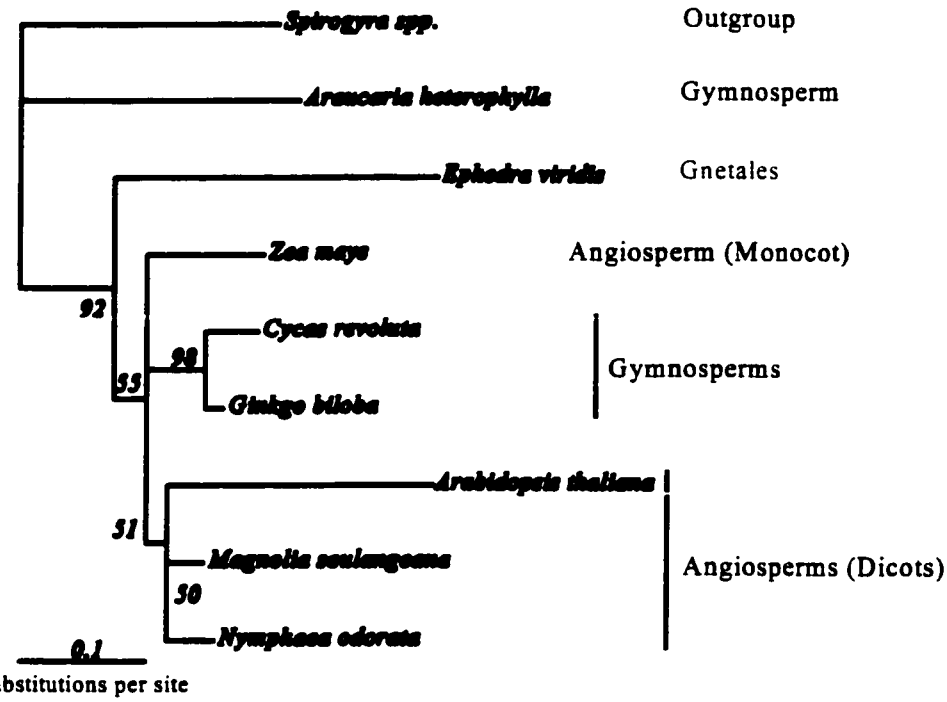


Figure 12A: Maximum-likelihood phylogeny inferred from a 280-position alignment of plant RPB1 D-F amino acid sequences using the Blosum62 model of substitution with an invariable rate and eight γ-distributed rates of substitution and maximum likelihood distances, using PUZZLE 4.0.2. The numbers at the nodes indicate the percent of 10000 quartet puzzling trees supporting that relationship, if >50%. The fraction of invariable sites is 0.41, the γ parameter α = 1.01 and log L = -1776.73 for the tree with nonclocklike branch lengths shown here.

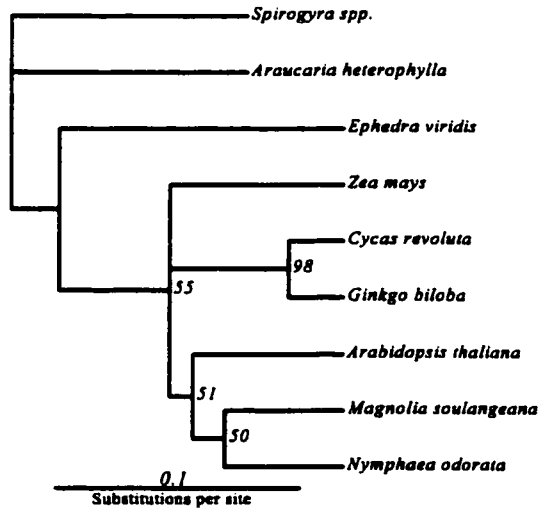


Fig.12B: Same as above, except with clocklike branch lengths. With log L = -1805.46, PUZZLE rejected this tree

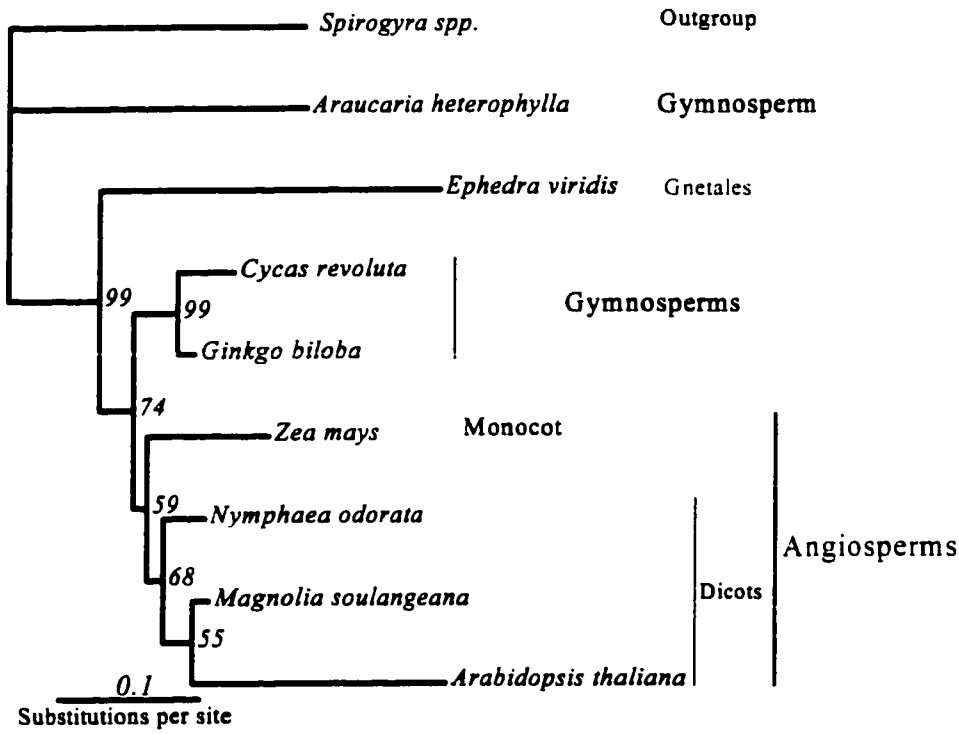


Figure 13A: A maximum likelihood distance phylogenetic tree inferred froma 280-position alignment of plant RPB1 D-F amino acid sequences using an invariable rate and eight γ-distributed rates of substitution according to the JTT model. Numbers at the nodes represent the percent of 10000 quartet puzzling steps supporting that relationship, if >50%. The frequency of invariable sites is 0.41, the γ parameter α = 1.01 and log L = -1754.94 for the nonclocklike branchlengths shown here.

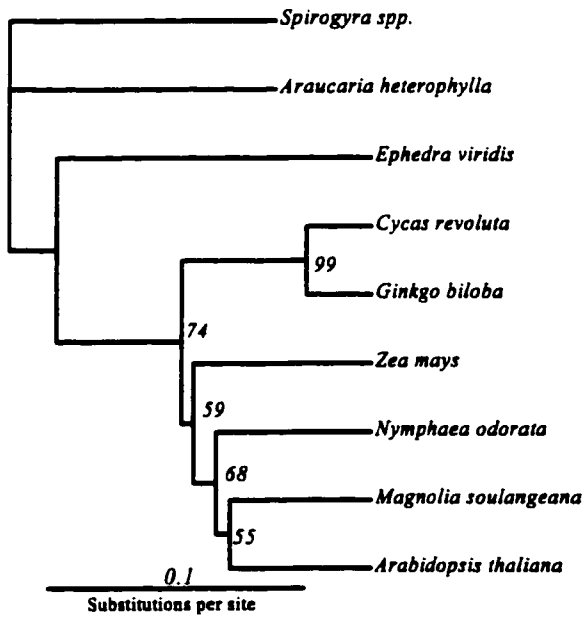


Fig. 13B: A tree with clocklike branch lengths, inferred using the JTT model and otherwise as described above. PUZZLE rejected this tree with log L = -1789.17.

RPB1 in other eukaryotes:

Figures 14, 15 and 16 represent phylogenetic trees of 19 RPB1 D-F amino acid sequences in animals, plants and fungi inferred by parsimony, neighbor-joining and maximum-likelihood methods for contrast with Figure 3, which used regions F-H of RPB1 from only 13 sequences, with another outgroup. The 282-position alignment represented by boxes in dotted lines in Figure 9 was used for this analysis, with the red alga *Porphyra* as the outgroup (see Appendix 3).

In Figure 14, only one of three equally parsimonious trees is shown since they only differed in their groupings of angiosperms amongst one another, none of which was supported by >50% bootstrap support. However, angiosperms were consistently grouped together exclusively, with the *Ginkgo* and *Cycad* clade as a sister group, and Gnetales as a sister group to that clade. However, *Araucaria*, which is a gymnosperm, was put in a more basal position than the green alga *Spirogyra* amongst plants. In the animal clade, both arthropods (*Drosophila* and *Artemia*) were grouped together with <50% bootstrap support, and both molluscs (*Crassostrea* and *Ilyanassa*) were grouped together, with bootstrap support. The nematode (*C. elegans*) was grouped as a sister to the arthropods with strong bootstrap support, and this clade was then grouped as a sister to the molluscs with <50% bootstrap support. The leech (*Helobdella*) was found basal to this group. The mouse represented all vertebrates and was placed as a sister group to all of the invertebrates mentioned above, with 100% bootstrap support. While RPB1 sequences from other vertebrates are available in Genbank, they are also both from placental mammals (human and Chinese hamster) and are identical to mouse at the amino acid level in domains D-F of RPB1, thus they were not included in this analysis. Animals and plants were related as sister groups, in contrast to figure 3.

The neighbor-joining analysis in Figure 15 also grouped animals and plants together. The topology of the plant branches of the tree are consistent with the neighbor-joining analysis in Figure 11, except that *Spirogyra* is placed as an ingroup in the second-most basal position amongst land plants, consistent with the parsimony results shown in Figure 14. The topology of the animal branches of the tree are consistent with Figure 3 only in the relationships amongst arthropods and molluscs. In contrast to the parsimony tree(s), *C. elegans* is placed as a sister group to the arthropod-mollusc clade, albeit with <50% bootstrap support. In Figure 3, *C.elegans* forms a clade with *Helobdella* and together these are a sister group to the arthropod-mollusc clade. However, mouse (*Mus*) is placed between *C.elegans* and *Helobdella* in Figure 15, making *Helobdella* the most basal animal clade, while mouse took that position in Figures 14 and 3. Fungi are grouped together, near the outgroup.

Figure 16 represents the tree with non-clocklike branch lengths that was obtained using the JTT model of maximum-likelihood analysis in PUZZLE. Animals and plants are related as sister groups in this tree. The relationships amongst the major groups of animals and plants, however, are largely unresolved. Amongst animals, arthropods are grouped together and molluscs are grouped together, but no relationships are made with these or the leech, nematode or vertebrate. Amongst plants, *Araucaria* and *Spirogyra* are grouped together and placed as the most basal clade of plants, sister to a mixed clade of angiosperms and gymnosperms. In this group, *Nymphaea* and *Magnolia* (two angiosperms) are grouped together, and *Cycas* and *Ginkgo* (two gymnosperms) are grouped together, but the relationship of these two groups with one another or with the remaining angiosperms and Gnetales are not resolved, in contrast with figures 12 and 13.

Figures 17 and 18 represent phylogenetic trees of 38 sequences of regions D-F of the largest subunit of RNA polymerases I, II and III (RPA1, RPB1 and RPC1) from a wider assortment of eukaryotes, inferred using the JTT and Blosum62 maximum likelihood methods in PUZZLE. The 252-position alignment represented by boxes in solid lines in Figure 9 was used for this analysis, with the Crenarchaeote archaeon *Sulfolobus acidocaldarius* RPA' as the outgroup (see Appendix 4). While this is a very distant outgroup, it was chosen in order to more clearly determine whether RPA1 or RPC1 are the sister group to RPB1. Both trees place RPA1 as the sister group to RPB1. In Figure 17, the relationships of the major eukaryotic groups to one another are not resolved by either RPB1 or by RPC1 sequences by the JTT model of substitution. The relationships seen in RPB1 sequences within animal taxa and within plants are similar to those in Figure 16, except that *Spirogyra* and *Araucaria* are not separated from other plants here. Other RPB1 clades seen in figure 17 are: red algae, fungi, Microsporidia, slime molds (*Dictyostelium* and *Acanthamoeba*) and a group with an Apicomplexan and a Parabasalid (*Plasmodium* and *Trichomonas*). In the RPA1 branch of the tree, animals and fungi are related as sister groups, in the absence of the newly published *Arabidopsis* RPA1 sequence.

The relationships seen in Figure 18 are similar to those in Figure 17, with a few notable exceptions. The relationship of *Ginkgo* and *Cycas* is lost in Figure 18, and the slime molds form a sister group to all plants. A Kinetoplastid - Archamoeba clade is formed with *Mastigamoeba* and *Trypanosoma* RPB1. Within the RPC1 sequences, a fungus and a Diplomonad (*Giardia*) form a clade, with a Kinetoplastid (*Trypanosoma*) as a sister group. Other relationships are essentially the same as those seen in Figure 17.

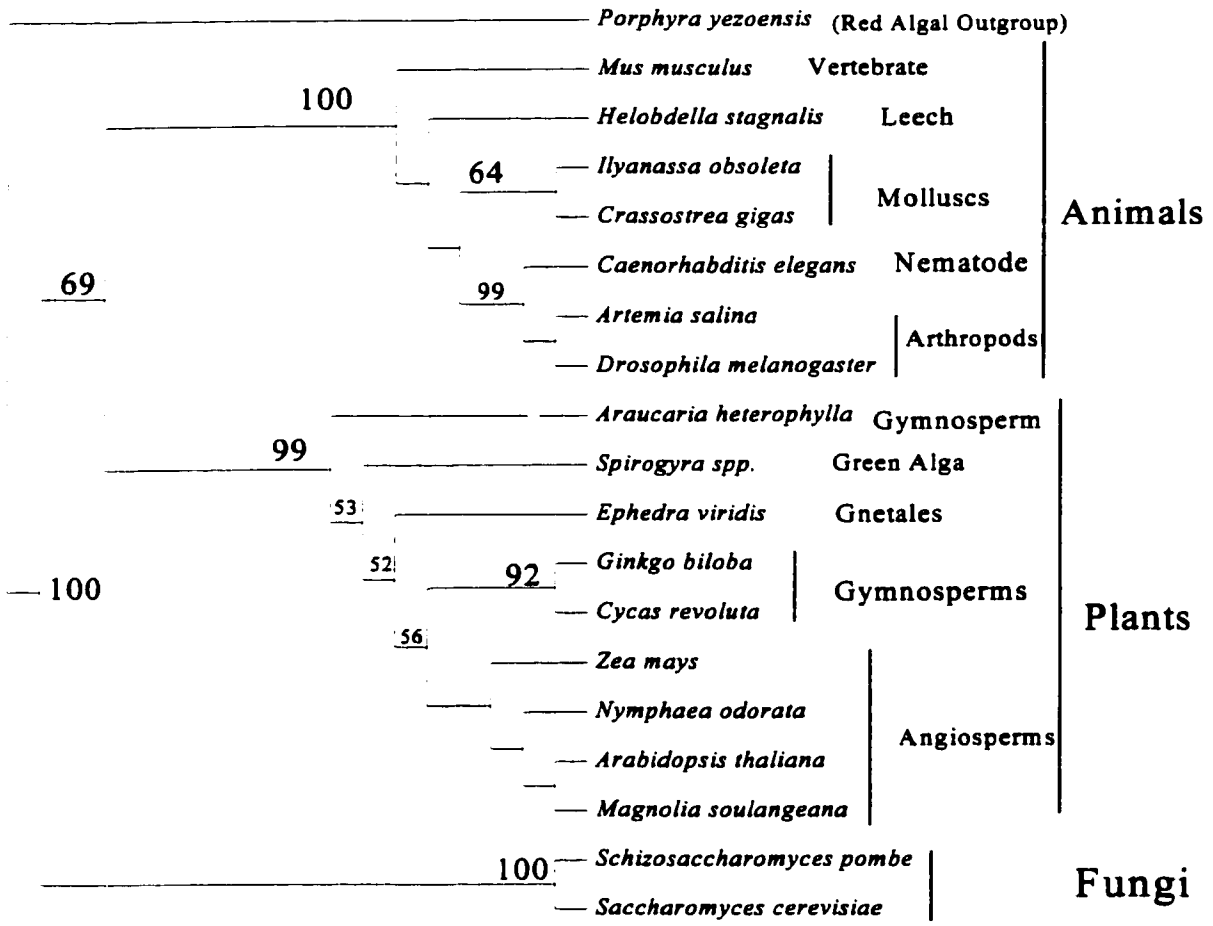


Figure 14: One of three equally most parsimonious phylogenetic trees inferred from a 282-position alignment of amino acid sequences of domains D-F of RPB1 from Animals, Plants and Fungi, using PROTPARS of PHYLIP 3.573c. The topology matching that of the consensus bootstrapped tree is shown, with the percent of 1000 bootstrap replicates supporting that topology shown at the nodes, if above 50%. *Porphyra*, a red alga, is the outgroup. The other two topologies differed only in that, amongst angiosperms, *Arabidopsis* remained in the same position, paired with either *Nymphaea* or *Zea*.

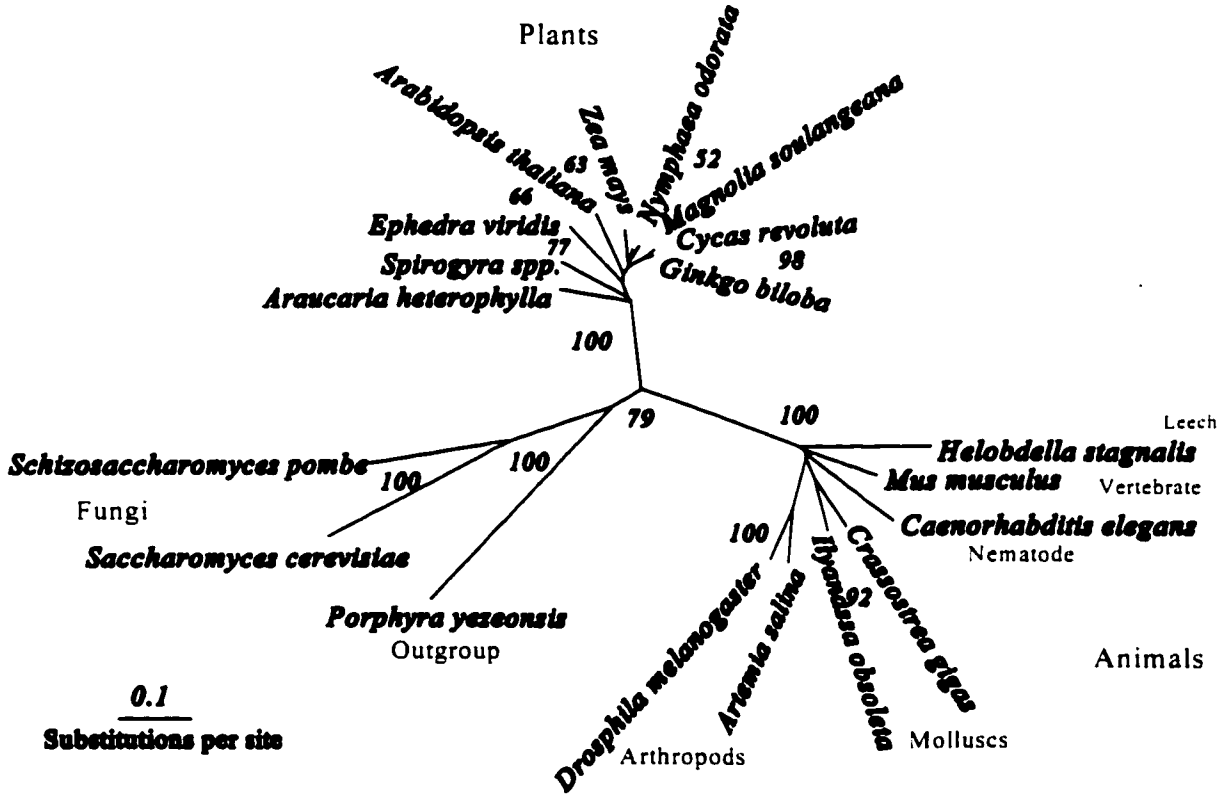


Figure 15: Neighbor-joining phylogenetic tree inferred from a 282-position alignment of RPB1 D-F amino acid sequences from Animals, Plants and Fungi, with a red alga as the outgroup, and distances inferred using a Dayhoff PAM distance matrix, using PROTDIST and NEIGHBOR in PHYLIP 3.573c. Numbers at the nodes represent the percent of 1000 bootstrap replicates supporting that relationship, if above 50%.

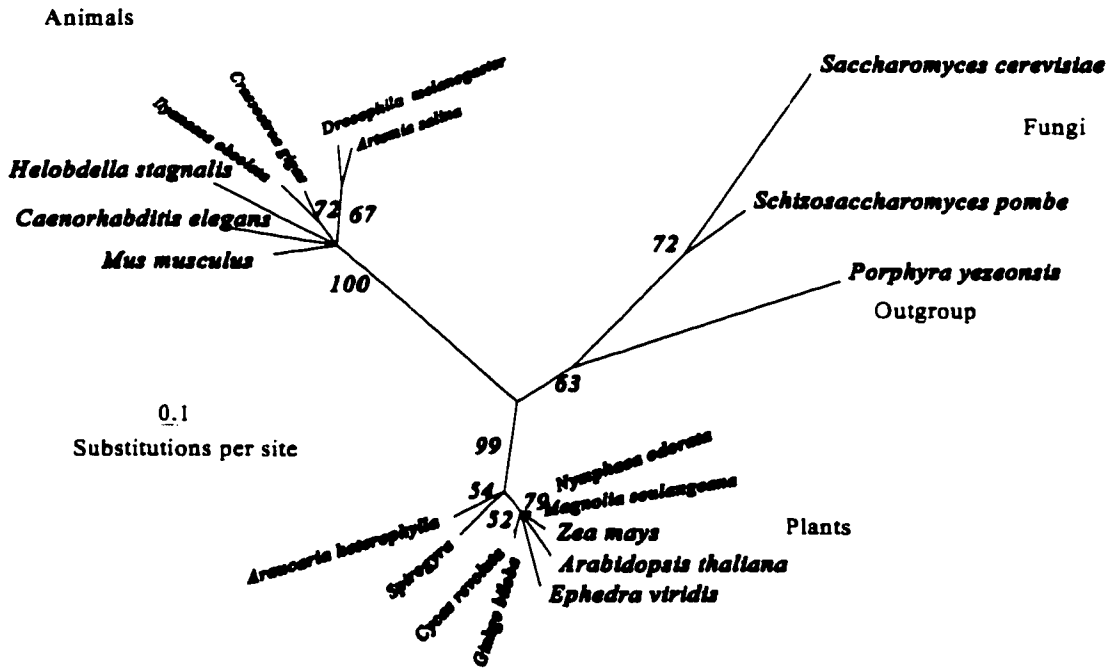


Figure 16: Maximum-likelihood distance phylogenetic tree with non-clocklike branch lengths of a 282-position alignment of RPB1 D-F amino acid sequences from Animals, Plants and Fungi, estimated using the JTT model of substitution and an invariable and eight γ-distributed rates of substitution with PUZZLE v.4.0.2. Numbers at the nodes represent the percent of 10000 quartet puzzling steps supporting that relationship, if >50%. The outgroup is a red alga. The γ parameter α = 0.87 and the fraction of invariable sites is 0.06. The log L of the tree shown here with non-clocklike branch lengths is -4849.98. The tree with clocklike branch lengths had a log L of -4897.91, and was rejected by PUZZLE and considered to be unlikely.

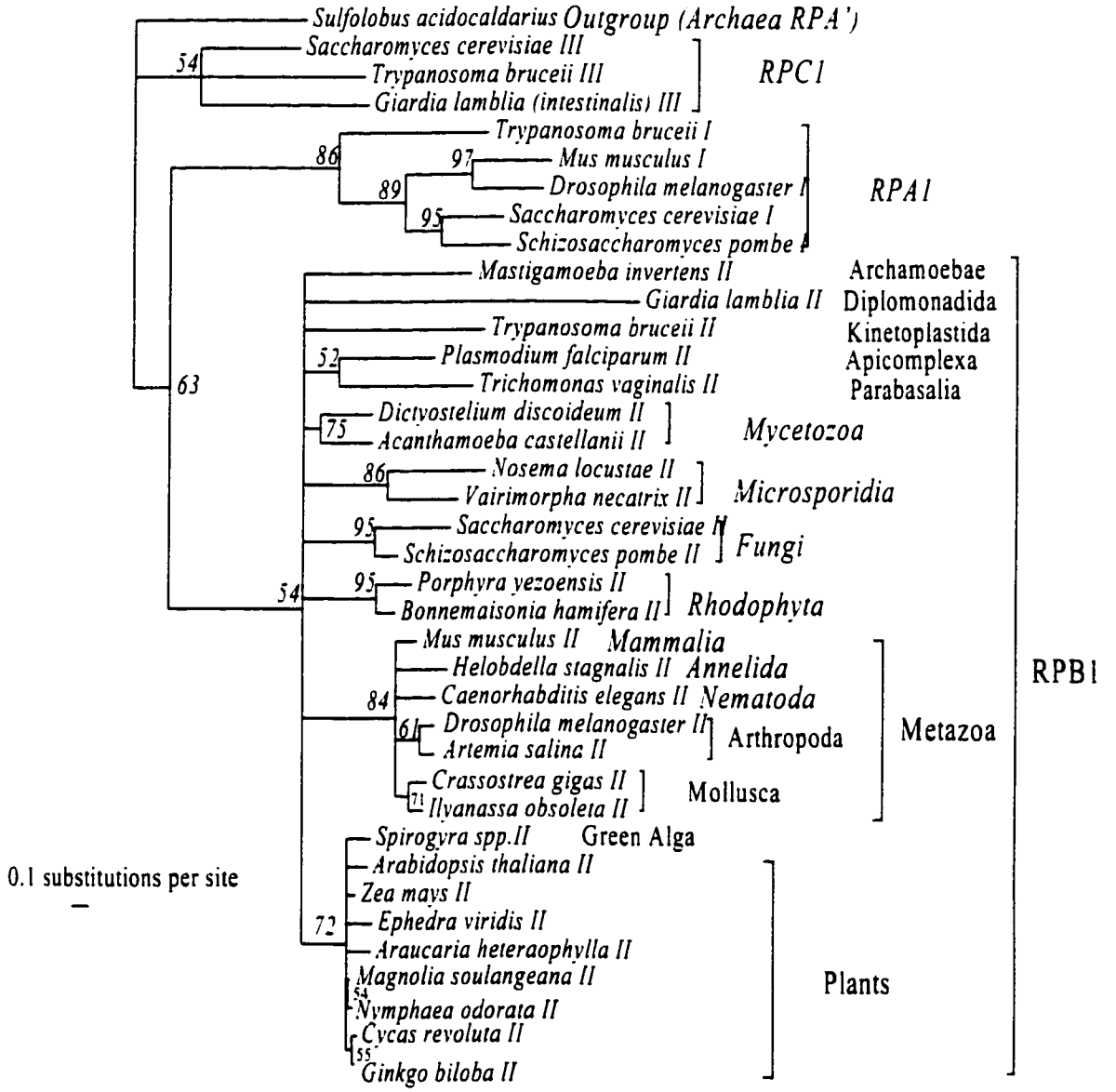


Figure 17: A maximum-likelihood distance phylogenetic tree inferred from a 252-position alignment of the amino acid sequences for domains D-F of the largest subunit of all eukaryotic RNA polymerases, inferred using the JTT model of substitution and considering an invariable and 8 γ-distributed rates of substitution, with the PUZZLE 4.0.2 program. The numbers at the nodes represent the percent of 10000 quartet puzzling steps supporting that topology, if above 50%. The fraction of invariable sites is 0.07 (S.E. 0.02) and γ parameter α = 1.72 (S.E. 0.18). The log L of th tree with non-clocklike branchlengths, shown here, is -12657.44, while the log L of the tree with clocklike branchlengths was -12788.63 and rejected by PUZZLE.

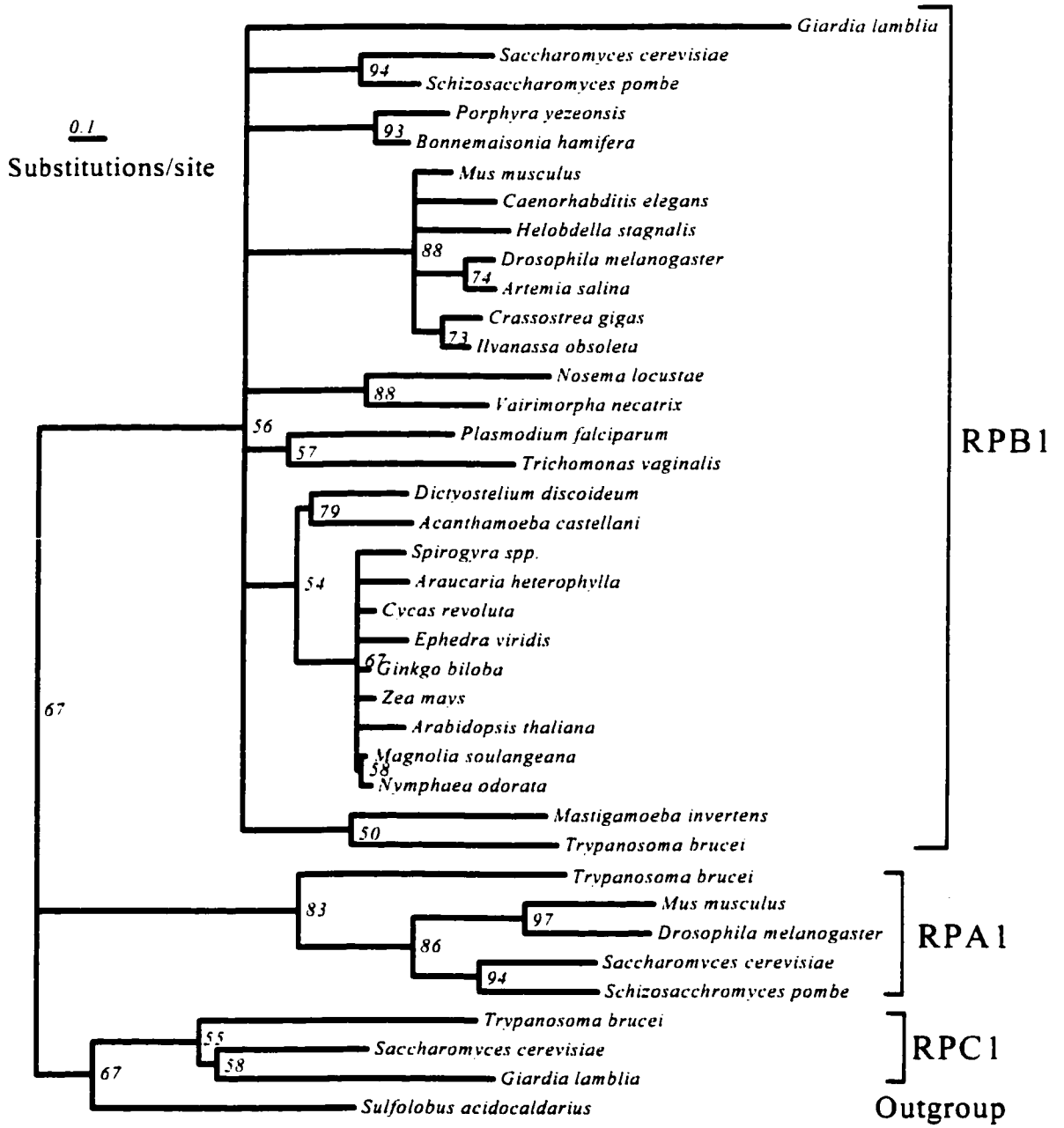


Figure 18: Maximum-likelihood distance phylogenetic tree inferred from a 252 residue alignment of amino acid sequences of domains D-F of the largest subunit of RNA polymerases I, II and III from all eukaryotes, with the archaeal homologue (RPA') from *Sulfolobus acidocaldarius* as the outgroup. The Blosum62 model of substitution was used in Puzzle 4.0.2, also considering an invariable rate as well as 8 gamma-distributed rates of substitution. Gaps and positions which could not be unambiguously aligned in more than three taxa were not included in the analysis. Numbers at the nodes represent the percent of 10000 quartet puzzling steps supporting that relationship, if over 50%. The log L for this tree with nonclocklike branch lengths is -12580.61 and the clocklike tree was rejected. The frequency of invariable sites is 0.07 and the γ distribution parameter $\alpha = 2.23$.

Compositional Bias:

The G+C content of the region of the RPB1 gene encoding domains D-F varies from 41% to 54% in plants and from 36% to 60% in a survey including most other eukaryotes (Figure 19). In plants and in other eukaryotes, this compositional bias is seen the most at the third base of codons, and at the first base of codons to a lesser degree. The nucleotide composition at the second base of codons is not affected by compositional bias. The compositional bias seen at the first base of codons is mostly due to bias in the first base of the sixfold degenerate arginine and leucine codons.

The analysis of G+C content at the third codon position of RPB1 regions D-F shown in Figure 20 illustrates the codon usage patterns seen amongst plants and other eukaryotes. For example, Figure 20B shows that most animals, most plants, and both fungi each share similar codon usage patterns. Figure 21 also shows the diversity of G+C *vs* A+T content at the third base of codons amongst plants for this region of RPB1. Within plants, the four angiosperms also share more similar nucleotide composition together than they do with the other plants. Fungi are 30-35% G+C at the third codon position, while most plants are 35-44% G+C and most animals fell into a broader range of 44-70% G+C, at the same codon position.

Figure 22 illustrates the effects of G+C compositional bias on amino acid usage. The G+C content in the coding region is compared to the content of amino acids with G+C rich codons (G, A, R, P) and to the content of amino acids with A+T rich codons (F, Y, M, I, N, K). Whether closely examining plants alone, or looking at a more broad range of eukaryotes, there appears to be little effect on amino acid usage from nucleotide compositional bias.

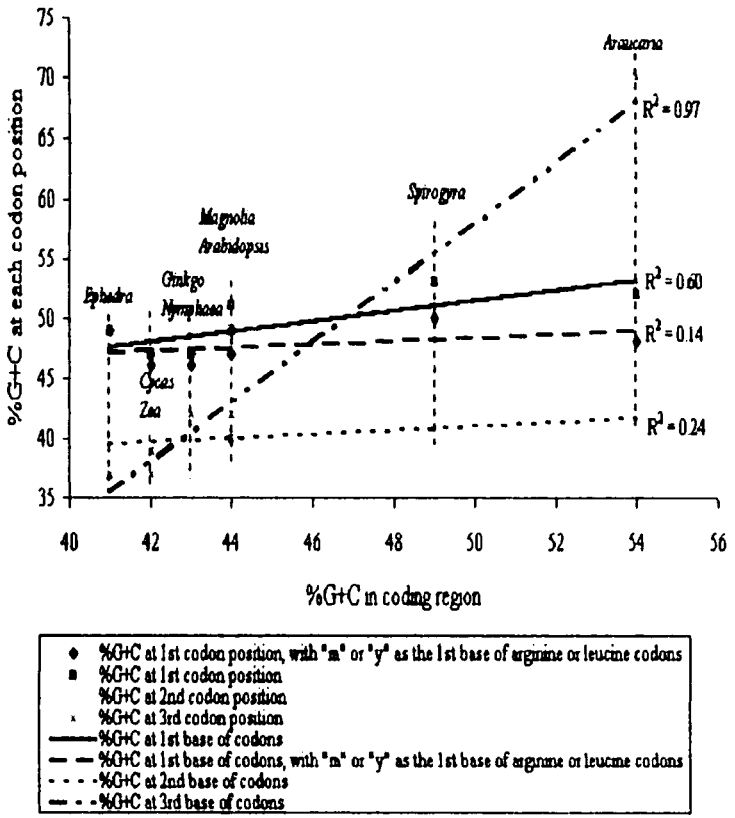


Figure 19. % G+C in the coding region vs. % G+C at each codon position of RPB1 D-F in plants, demonstrating the effects of compositional bias at the first and third bases of codons

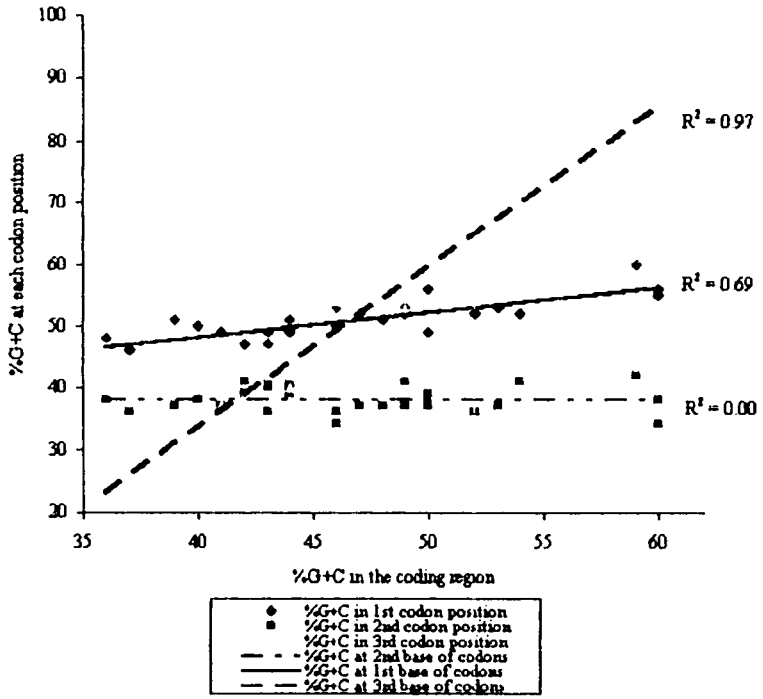


Figure 19B The effect of G+C content on each codon position of RPB1 D-F in eukaryotes

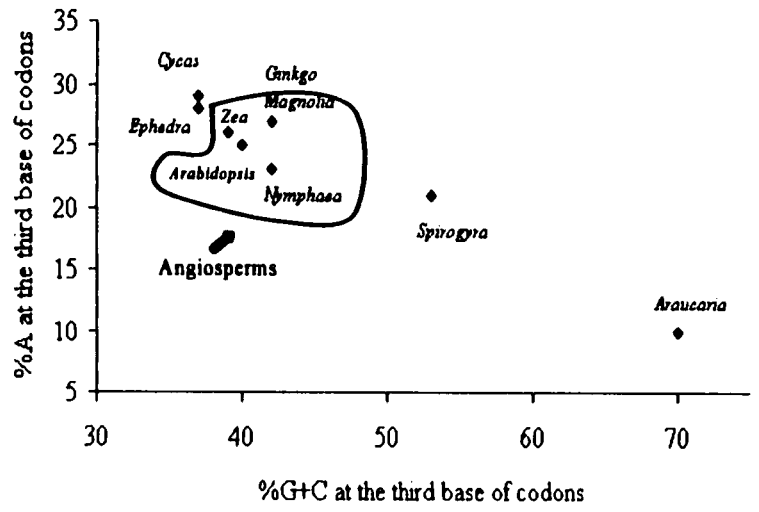
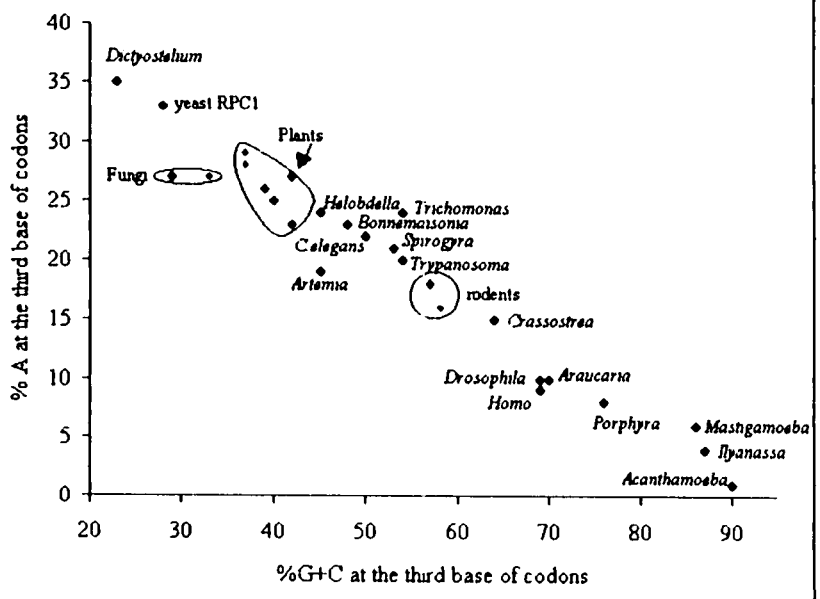


Figure 20: %G+C vs %A at the third base of codons of RPB1 D-F in plants

Figure 20B. G+C content at the third base of codons reflects codon usage patterns in RPB1 D-F amongst eukaryotes



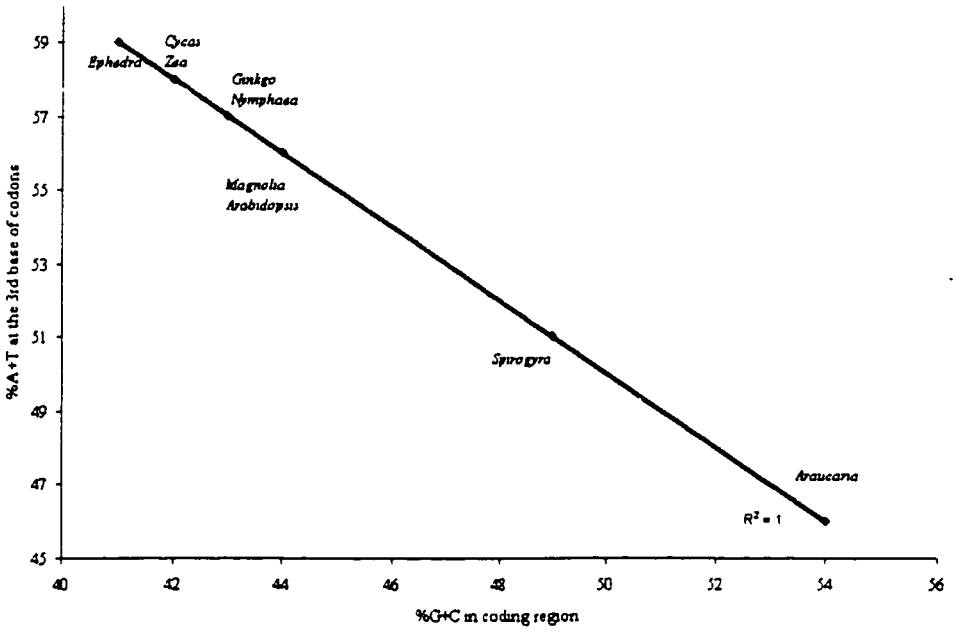


Figure 21: %G+C of coding region vs % A+T at the third codon position in RPB1 D-F of plants, demonstrating the variation in G+C content among plant RPB1 D-F sequences.

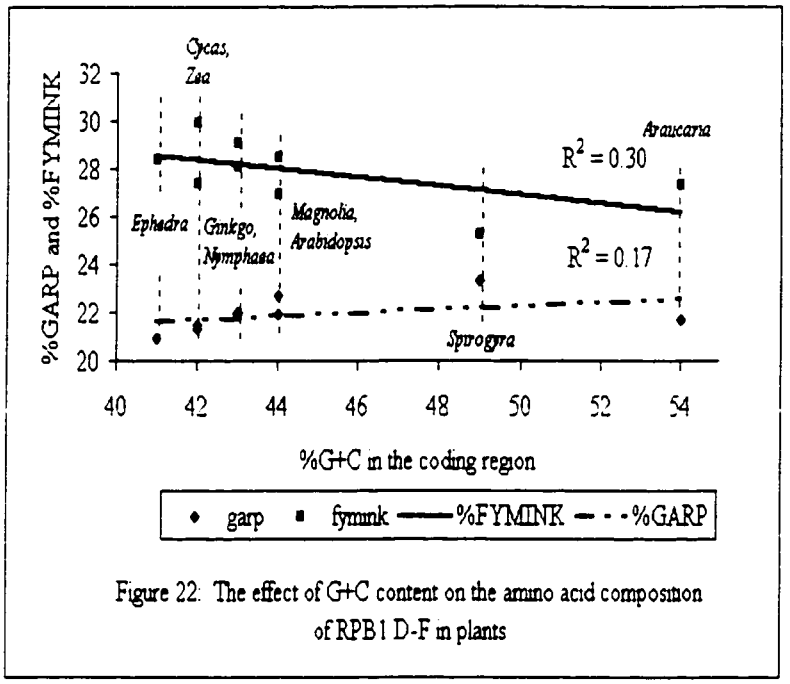


Figure 22: The effect of G+C content on the amino acid composition of RPB1 D-F in plants

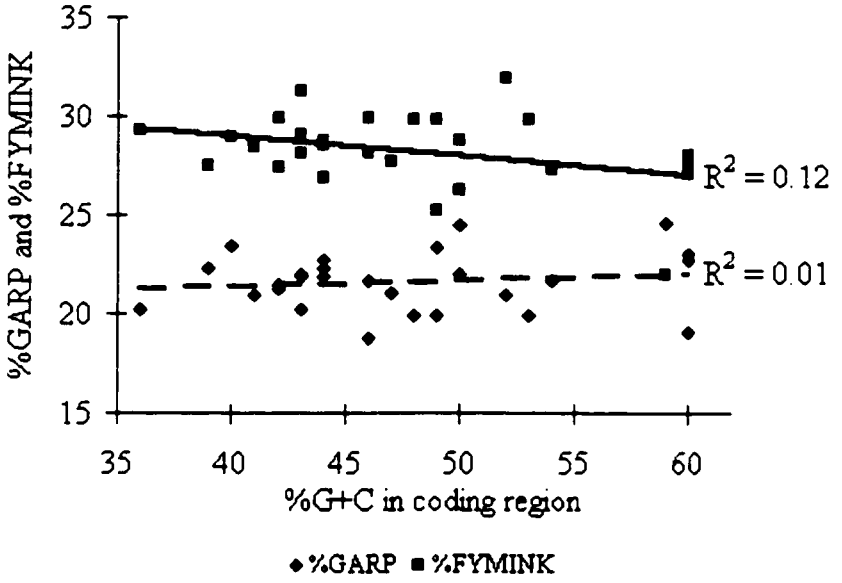


Figure 22B: Effect of G+C content on the amino acid composition of RPB1 D-F in eukaryotes

Substitution Analysis of Region D-F of RPB1 in Land Plants:

Tables 5 and 6 show the average nucleotide substitutions per site amongst the 834 bp alignment of plant sequences shown in figure 5, as calculated using the Li93 program. Table 5 shows the average nonsynonymous substitutions per site (Ka), when they were calculated to be less than 3.24 by the Li93 method. According to the Jukes-Cantor formula $K = -\frac{3}{4} \ln(1-\frac{4}{3}P)$, if the number of substitutions observed (P) is ≥ 0.75, then K is undefined. In this situation, Li93 cannot calculate K. Those numbers greater than 3.24 result from the presence of too many multiple substitutions. The Jukes-Cantor formula was used to estimate how many of the 556 nonsysnonymous sites in the alignment had multiple substitutions. This is summarized in Table 7. Table 6 shows the average synonymous substitutions per site when they were calculated to be less than 3.24 since other values represented the presence of too many multiple substitutions. The estimate of the number of 278 synonymous sites in the alignment that had multiple substitutions is summarized in Table 8. These results show that the synonymous sites are saturated with multiple substitutions, while the incidence of multiple substitutions at nonsynonymous sites is approximately 100 times less.

The Ka of *Nymphaea* and *Magnolia vs* maize and *vs Araucaria, Ginkgo* and *Cycas*, as well as the Ka of *Spirogyra vs Araucaria* and *Nymphaea* compared with the approximate divergence times of these groups is shown in Figure 23. Other comparisons are not shown in this graph if either the divergence time is uncertain or the Ka (see Table 5) is undefined. Major conclusions should not be made from this graph without the inclusion of more data from vascular plants at least.

Table 5: The weighted average of nonsynonymous substitutions per site (K_a), with standard error, calculated for the nine plant RPB1 D-F nucleotide sequences by Li93. The $P_a = ¾ (1\text{-}e^{-(4/3) Ka})$ is shown underneath in parentheses. Dashes represent undefined terms.

Standard Error

Nonsynonymous substitutions per site

	Spirogyra sp.	*Araucaria heterophylla*	*Cycas revoluta*	*Ephedra viridis*	*Ginkgo biloba*	*Magnolia soulangeana*	*Nymphaea odorata*	*Zea mays*	*Arabidopsis thaliana*
Spirogyra sp.		0.016	-	0.016	-	-	0.014	-	-
Araucaria heterophylla	0.136 (0.124)		-	-	-	0.013	-	-	-
Cycas revoluta	-	-		0.013	0.005	0.008	0.008	0.01	0.011
Ephedra viridis	0.137 (0.125)	-	0.097 (0.091)		0.013	0.013	0.013	0.013	0.014
Ginkgo biloba	-	-	0.018 (0.018)	0.093 (0.087)		0.007	0.007	0.01	0.012
Magnolia soulangeana	-	0.093 (0.087)	0.041 (0.040)	0.094 (0.088)	0.035 (0.034)		0.006	0.01	0.009
Nymphaea odorata	0.114 (0.106)	-	0.039 (0.038)	0.095 (0.089)	0.029 (0.028)	0.022 (0.022)		0.01	0.01
Zea mays	-	-	0.049 (0.047)	0.101 (0.094)	0.047 (0.046)	0.034 (0.033)	0.043 (0.042)		-
Arabidopsis thaliana	-	-	0.077 (0.073)	0.105 (0.098)	0.080 (0.076)	0.053 (0.051)	0.065 (0.062)	-	

Table 6: The weighted average of synonymous substitutions per site (K_s), with standard error, calculated for the nine plant RPB1 D-F nucleotide sequences by Li93. P_s is shown underneath in parentheses. Dashes represent undefined terms.

Standard Error

Synonymous substitutions per site (Ks)

	Spirogyra sp.	*Araucaria heterophylla*	*Cycas revoluta*	*Ephedra viridis*	*Ginkgo biloba*	*Magnolia soulangeana*	*Nymphaea odorata*	*Zea mays*	*Arabidopsis thaliana*
Spirogyra sp.		-	-	-	-	-	-	-	-
Araucaria heterophylla	-		-	-	-	-	-	-	-
Cycas revoluta	-	-		0.258	0.077	0.226	0.276	3.882	2.52
Ephedra viridis	-	-	1.323 (0.621)		0.261	0.824	-	-	-
Ginkgo biloba	-	-	0.562 (0.395)	1.451 (0.642)		0.267	0.68	-	-
Magnolia soulangeana	-	-	1.31 (0.62)	1.96 (0.70)	1.566 (0.657)		0.209	0.242	0.507
Nymphaea odorata	-	-	1.456 (0.642)	-	2.28 (0.71)	1.318 (0.621)		0.346	0.769
Zea mays	-	-	2.835 (0.733)	-	-	1.476 (0.645)	1.669 (0.669)		-
Arabidopsis thaliana	-	-	2.988 (0.736)	-	-	1.889 (0.690)	2.162 (0.708)	-	

Table 7: Estimate of the number of multiple substitutions at nonsynonymous (n.s.) sites in plant RPB1 D-F sequences shown in Figure 5, calculated using the difference between the K_a and P_a values shown in Table 5.

species 1	species 2	Inferred frequency of n.s. substitutions (K_a)	Observed frequency of n.s. substitutions (P_a)	% multiple substitutions at n.s. sites	# of multiple substitutions in 556 n.s. sites
Araucaria	*Spirogyra*	0.136	0.124	1.2 %	7
Ephedra	*Spirogyra*	0.137	0.125	1.2 %	7
Nymphaea	*Spirogyra*	0.114	0.106	0.8 %	4
Araucaria	*Magnolia*	0.093	0.087	0.6 %	3
Cycas	*Ephedra*	0.097	0.091	0.6 %	3
Cycas	*Ginkgo*	0.018	0.018	0 %	0
Cycas	*Magnolia*	0.041	0.04	0.1 %	1
Cycas	*Nymphaea*	0.039	0.038	0.1 %	1
Cycas	*Zea*	0.049	0.047	0.2 %	1
Cycas	*Arabidopsis*	0.077	0.073	0.4 %	2
Ephedra	*Ginkgo*	0.093	0.087	0.6 %	3
Ephedra	*Magnolia*	0.094	0.088	0.6 %	3
Ephedra	*Nymphaea*	0.095	0.089	0.6 %	3
Ephedra	*Zea*	0.101	0.094	0.7 %	4
Ephedra	*Arabidopsis*	0.105	0.098	0.7 %	4
Ginkgo	*Magnolia*	0.035	0.034	0.1 %	1
Ginkgo	*Nymphaea*	0.029	0.028	0.1 %	1
Ginkgo	*Zea*	0.047	0.046	0.1 %	1
Ginkgo	*Arabidopsis*	0.08	0.076	0.4 %	2
Magnolia	*Nymphaea*	0.022	0.022	0 %	0
Magnolia	*Zea*	0.034	0.033	0.1 %	1
Magnolia	*Arabidopsis*	0.053	0.051	0.2 %	1
Nymphaea	*Zea*	0.043	0.042	0.1 %	1
Nymphaea	*Arabidopsis*	0.065	0.062	0.3 %	2

Table 8: Estimate of the number of multiple substitutions at synonymous (syn.) sites in plant RPB1 D-F sequences shown in Figure 5, calculated using the difference between the K_s and P_s values shown in Table 6.

species 1	species 2	Inferred frequency of syn. substitutions (K_s)	Observed frequency of syn. substitutions (P_s)	% multiple substitutions at syn. sites	# of multiple substitutions in 278 syn. sites
Cycas	*Ephedra*	1.323	0.621	70.2 %	195
Cycas	*Ginkgo*	0.562	0.395	16.7 %	46
Cycas	*Magnolia*	1.31	0.62	69 %	192
Cycas	*Nymphaea*	1.456	0.642	81.4 %	226
Cycas	*Zea*	2.835	0.733	210.2 %	584
Cycas	*Arabidopsis*	2.988	0.736	225.2 %	626
Ephedra	*Ginkgo*	1.451	0.642	80.9 %	225
Ephedra	*Magnolia*	1.96	0.7	126 %	350
Ginkgo	*Magnolia*	1.566	0.657	90.9 %	253
Ginkgo	*Nymphaea*	2.28	0.71	157 %	436
Magnolia	*Nymphaea*	1.318	0.621	69.7 %	194
Magnolia	*Zea*	1.476	0.645	83.1 %	231
Magnolia	*Arabidopsis*	1.889	0.69	119.9 %	333
Nymphaea	*Zea*	1.669	0.669	100 %	278
Nymphaea	*Arabidopsis*	2.162	0.708	145.4 %	404

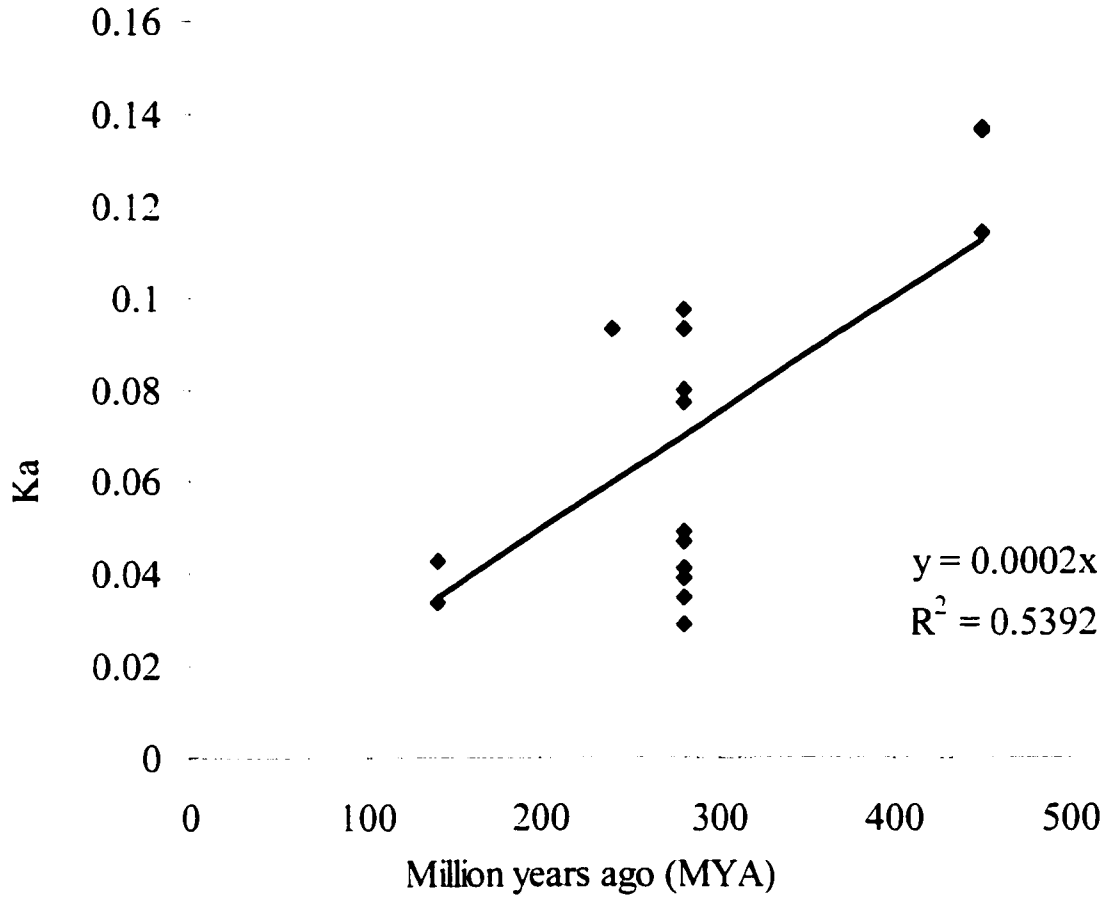


Figure 23: Divergence times vs nonsynonymous substitutions per site (Ka) in plant RPB1D-F sequences. Ka values (see Table 7) of *Magnolia* and *Nymphaea vs* maize, *Cycas* and *Ginkgo*, of *Araucaria vs Spirogyra* and *Magnolia*, of *Spirogyra vs Nymphaea* and *Ephedra*, and of maize, *Ephedra* and *Arabidopsis vs Cycas* and *Ginkgo* were compared to divergence times adapted from Crane *et al* (1995), Cronquist (1981) and Stewart and Rothwell (1993). Some comparisons (e.g., *Magnolia vs Nymphaea)* were not plotted because of unknown divergence times, while others were not plotted because the Ka was undefined.

Relative Rate Tests:

Relative rate tests were done using the RRTree program with amino acid sequence alignments of plants, of animals, plants and fungi, and of RPA1, RPB1 and RPC1. These results are summarized in Tables 9-12. Each plant sequence was compared individually (Table 9), while in the other analyses, groups of sequences were compared to each other using the group's mean substitution rate (K), with each taxon in the group equally weighted (Tables 10-12). The p-value for the exact probability that the two groups have the same substitution rates was tested against $\alpha = 0.05/(\text{number of comparisons})$. No significantly different rates of substitution were found.

Table 9: Results of relative rate tests of substitution rates (K) of amino acid sequences of RPB1 regions D-F amongst plants, based upon the alignments shown in Figure 9 and also in Appendix 2. Since 28 pairwise comparisons were made in this analysis, $\alpha = 0.05/28 = 0.002$. *Spirogyra* is the outgroup. P-values are given in the upper upper corner of the table, while the difference in substitution rate is given in the bottom of the table.

p-value

difference in substitution rate (ΔK)

	Araucaria heterophylla	*Cycas revoluta*	*Ephedra viridis*	*Ginkgo biloba*	*Magnolia soulangeana*	*Nymphaea odorata*	*Zea mays*	*Arabidopsis thaliana*
Araucaria heterophylla		0.765	0.895	0.678	0.715	0.833	0.891	0.894
Cycas revoluta	0.04		0.611	0.809	0.958	0.874	0.804	0.611
Ephedra viridis	-0.018	-0.058		0.542	0.607	0.704	0.763	0
Ginkgo biloba	0.053	0.013	0.071		0.909	0.717	0.689	0.542
Magnolia soulangeana	0.044	0.004	0.062	-0.009		0.767	0.729	0.516
Nymphaea odorata	0.027	-0.013	0.045	-0.026	-0.017		0.914	0.668
Zea mays	0.018	-0.022	0.036	-0.035	-0.026	-0.009		0.756
Arabidopsis thaliana	-0.018	-0.058	0	-0.071	-0.062	-0.045	-0.036	

Table 10: Results of relative rate tests of substitution rates (K) of amino acid sequences of RPB1 regions D-F between animals, plants and fungi, based upon the alignments shown in Figure 9 and also in Appendix 3. Since three pairwise comparisons were made in this analysis, $\alpha = 0.05/3 = 0.017$. *Porphyra* is the outgroup.

Group 1	Group 2	ΔK	p-value
animals	plants	-0.136	0.397
animals	fungi	-0.027	0.878
plants	fungi	0.108	0.526

Table 11: Results of relative rate tests of substitution rates (K) of amino acid sequences of RPB1 regions F-H between animals, plants and fungi, based upon the alignment shown in Figure 2. Since three pairwise comparisons were made in this analysis, $\alpha = 0.05/3 = 0.017$. *Dictyostelium* is the outgroup.

Group 1	Group 2	ΔK	p-value
animals	plants	0.177	0.264
animals	fungi	-0.065	0.642
plants	fungi	0.242	0.142

Table 12: Results of relative rate tests of substitution rates (K) of amino acid sequences of regions D-F of RPA1, RPB1 and RPC1 in eukaryotes, based upon the alignments shown in figures 9 and in Appendix 4. Since three pairwise comparisons were made in this analysis, $\alpha = 0.05/3 = 0.017$. *Sulfolobus* is the outgroup.

Group 1	Group 2	ΔK	p-value
RPA1	RPB1	-0.221	0.34
RPA1	RPC1	-0.268	0.253
RPB1	RPC1	0.048	0.825

Copy Number of Plant RPB1s:

Southern analysis of *Magnolia* and *Nymphaea* genomic DNA probed with homologous RPB1 D-F clones by G. Drouin showed the presence of multiple signals. The presence of hybridization signals 7-11 kb in size verifies the source of the cloned RPB1 PCR products. The largest multiple bands indicate that multiple copies of RPB1 genes exist in these genomes. The presence of several smaller bands in lanes 1 and 3 of Figure 24A and in lane 2 of Figure 24B may indicate that some of the copies of RPB1 were cut internally. Table 13 shows the ploidy of the plants used in this study. Both *Magnolia soulangeana* and *Nymphaea odorata* are polyploid, with chromosome numbers of 6n = 114 and 12n = 84, respectively. It is possible that the multiple copies of RPB1 are a result of this polyploidy and/or gene duplication(s). Since the blots were washed under moderate stringency, the degree of identity of the hybridization signals with the probe is not known, so RPB1 pseudogenes may have hybridized if they were present. These two Southern blots could not specifically rule out the possibility that paralogous RPB1 sequences from *Magnolia soulangeana* and *Nymphaea odorata* were being used for phylogenetic analysis.

Initial hybridizations with the Northern blots (by B.M.) did not work (results not shown) and were not repeated, but the blots have been stored for future use. Southern blots of *Cycas* and *Ginkgo* (by B.M., results not shown) total DNA digested with *Eco*RI, *Bam*HI, or a double digest of both enzymes, and probed with homologous probes of RPB1 D-F cloned PCR products revealed hybridization of 8-kb bands, which could contain the entire coding region and intron sequences of RPB1. However, these membranes had a lot of background hybridization and it is difficult to confirm or rule out the presence of smaller bands such as those seen in Figure 24. Southern analysis of *Araucaria* total DNA did not work as a result of prevailing difficulty in obtaining sufficiently

clean DNA for electrophoresis of a small, concentrated sample volume. Southern blots were prepared for *Osmunda, Psilotum, Marchantia* and *Equisetum* but have not been hybridized and are stored for later use. Southern analysis of *Ephedra* and maize did not work and needs to be repeated.

It is important to note three points in favor of orthology of other RPB1 sequences used in this analysis. First of all, RPB1 is known to exist as a single copy in all eukaryotes besides *Trypanosoma brucei* and *Glycine max*, although this has not been investigated further in diploid RPB1 sequences cloned in this study. The expressed genes from *Arabidopsis thaliana, Zea mays,* and *Spirogyra* were used for the phylogenetic analysis of RPB1 in plants. The three intron positions and the length of the coding sequence in regions D-F found previously in *Arabidopsis* and in *Spirogyra* were perfectly conserved in all of the plant RPB1 sequences.

Table 13 identifies several (*) species from which RPB1 PCR products were not cloned. Prevailing difficulty was experienced in isolating sufficiently clean and abundant samples of DNA from *Osmunda, Equisetum, Marchantia* and *Psilotum,* but most of all with *Osmunda*. Cloning RPB1 D-F was also complicated by the PCR amplification of multiple PCR products, which probably indicate gene duplications compounded by the ancient polyploidy (Stebbins, 1950) of these species. RPB1-like clones were never obtained from *Gnetum* and *Welwitschia.*

Figure 24 A and B: Southern blots of *Magnolia soulangeana* (A) and *Nymphaea odorata* (B) total genomic DNA digests probed with homologous clones of regions D-F of RPB1. Lanes 1, 2 and 3 represent single digests with *Eco*RI, *Pst*I and *Xba*I. The λ*Bst*EII size marker is in the centre.

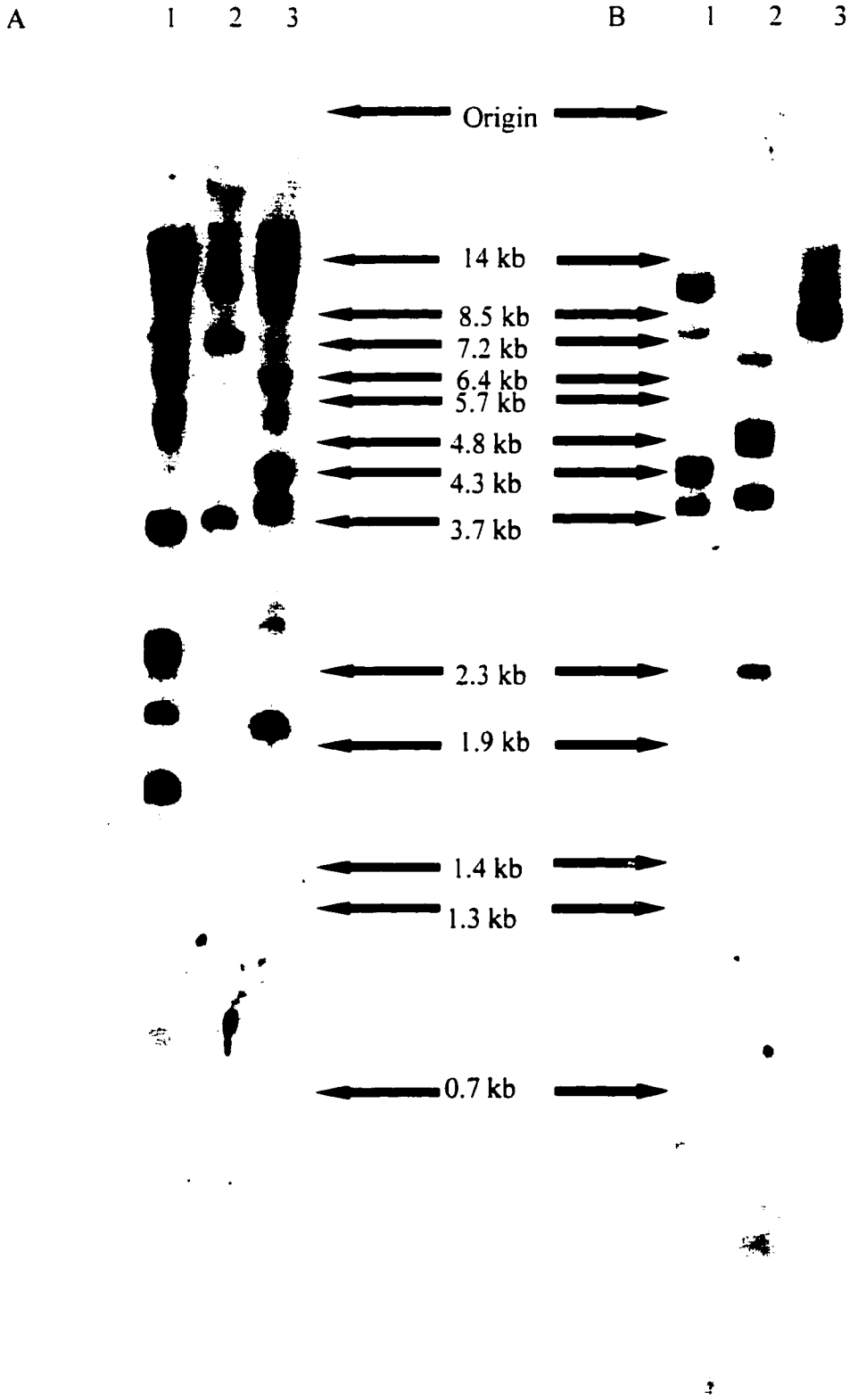


Table 13: Ploidy of representatives of some major plant phyla (Bennett *et.al.*, 1982; Bennet and Smith, 1976; Darlington and Ammal, 1945; Sparrow *et.al.*, 1972; Stebbins, 1950).

SPECIMEN	CHROMOSOME NUMBER
Arabidopsis thaliana	n=5 (2n)
Araucaria heterophylla	n=13 (those species listed were all 2n)
Cycas revoluta	n=11,12 (2n)
Ephedra viridis	n=7 (2n or 4n)
*Equisetum arvense**	n=108, polyploid
Ginkgo biloba	n=12 (2n)
*Gnetum gnemon**	n=? (those *Gnetum* species listed were 2n=24)
Magnolia soulangeana	n=19 (6n)
*Marchantia polymorpha**	n=9, ploidy uncertain
Nymphaea odorata	n=7, 14 (12n=84)
*Osmunda cinnamomea**	n=22, likely polyploid
*Psilotum nudum**	n=104, polyploid
*Welwitschia mirabilis **	n=7 (6n or 12n found)
Zea mays	n=10 (2n)

*species from which RPB1 was not cloned.

DISCUSSION

Phylogenetic Analysis of DNA Sequences of Regions D-F of RPB1 in Plants:

Only one relationship amongst plants was consistently resolved using three different methods of phylogenetic analysis of RPB1 D-F nucleotide sequences. *Cycas* and *Ginkgo* were always grouped together, with strong support. On the basis of this analysis, due to the conflicting phylogenies obtained using different methods to analyse the same dataset, neither the woody nor herbaceous origin of angiosperms can be supported. Similarly, the position of the Gnetales relative to gymnosperms and angiosperms is also unclear. The grouping of *Cycas* and *Ginkgo* is consistent with results found by Chaw *et al* (1997) with an 18S rRNA dataset and by Goremykin *et al* (1996) with an *rbc*L and chloroplast intergenic transcribed spacer (cpITS) sequence dataset.

The data about nucleotide compositional bias in plant RPB1 D-F sequences, described in figures 19 through 21, reveals that compositional bias likely had little effect upon this set of phylogenetic analyses. The trend of nucleotide composition of the coding region had negligible impact upon the second base of codons in this sequence, nor does it impact the first codon position after the neutral symbols "m" and "y" were substituted for the first base of arginine and leucine codons. It is also unlikely that the inconsistencies in the trees in figures 6-8 result from the similar codon usage patterns shared by seven of the eight ingroup plant taxa, which are simply a unifying characteristic of plant RPB1 D-F. The short length of this sequence (556 nonsynonymous sites) and the conserved nature of this part of the coding region of RPB1 probably did not provide enough informative variation to clearly infer the phylogeny of these plants (Nei, 1996).

codons. It is also unlikely that the inconsistencies in the trees in figures 6-8 result from the similar codon usage patterns shared by seven of the eight ingroup plant taxa, which are simply a unifying characteristic of plant RPB1 D-F. The short length of this sequence (556 nonsynonymous sites) and the conserved nature of this part of the coding region of RPB1 probably did not provide enough informative variation to clearly infer the phylogeny of these plants (Nei, 1996).

It also appears that the distance and parsimony trees constructed with the RPB1 nucleotide dataset were more prone to "long branch attraction" than was the PUZZLE maximum-likelihood (ML) method. With the former two methods, the most different sequences fell to the bottom of the trees. The latter method separated *Ephedra* and *Arabidopsis*, which both had long branches, away from the bottom of the tree (where both of the other long branches lie) and positioned them closer within the ingroups. However, even there, these two taxa were still grouped together, albeit with low support. The results shown in figure 8 which use the ML method and account for substitution rate heterogeneity in different parts of the alignment were the most robust results obtained from the nucleotide data. However, this method is not enough to resolve relationships in the absence of sufficient informative variation.

Phylogenetic Analysis of Amino Acid Sequences of Regions D-F of RPB1 in Plants:

Figure 13(A) represents the most robust tree obtained from analysis of RPB1 D-F in plants, since it inferred specific relationships of seven of the eight ingroups with one another, each with 55% or greater statistical support, unlike any of the previous trees in this study. This shows the difference made by using a method which accounts for substitution rate heterogeneity amongst sequences.

sequences by Goremykin *et al* (1996) and Hasebe (1992) and by different analyses by Hansen *et al* (1999) of a 9149 bp alignment which included several chloroplast genes and nuclear-encoded 18S rRNA. This relationship is also supported by the analysis of the MADS-box multigene family by Winter *et al* (1999) and by the analysis of a concatenated dataset of genes encoded in the nucleus, mitochondria and chloroplasts by Qiu *et al* (1999).

Figure 13 also weakly supports the herbaceous origin of dicots, but the position of monocots as the most basal angiosperms conflicts with most other molecular and morphological datasets. Parsimony analysis (figure 10 A and B) also weakly supports a herbaceous origin of angiosperms in two of four equally parsimonious trees, while neighbor-joining (figure 11) cannot resolve this relationship. RPB2 analysis (Denton *et al*, 1998) also supports a herbaceous origin of angiosperms but further comparisons of plant RPB2 and RPB1 analyses are limited by the different taxa used in each dataset. With *Hordeum* representing monocots in the RPB2 tree, rather than *Zea* in the RPB1 tree, RPB2 placed monocots amongst the woody dicots rather than at the base of the angiosperm tree. Recent analyses of sequences encoded in the nucleus, mitochondria and chloroplasts agree that angiosperms had a herbaceous origin, placing *Nymphaea* as one of the earliest angiosperms (Mathews and Donoghue, 1999; Parkinson *et al*, 1999; Qiu *et al*, 1999; Soltis *et al*, 1999; see Appendix 5 and discussion below).

As with the parsimony and neighbor-joining phylogenetic trees inferred from plant RPB1 nucleotide data, the parsimony and neighbor-joining amino acid trees also appear to show "long branch attraction". This flaw is shaken apart somewhat using the Blosum62 model of substitution in the maximum-likelihood analysis, which draws *Arabidopsis* amongst other angiosperms rather than stuck amongst gymnosperms. The JTT (Jones, Taylor, Thornton) model of substitution not

only does this, however, but also resolves more relationships from this dataset, with high statistical support.

The discrepancy between distance and parsimony methods is discussed in detail by Kuhner and Felsenstein (1994) as well as by Tateno *et al* (1994) concluding that the difference in substitution rates amongst different taxa can cause inaccuracy in parsimony analysis, and generally that in the case of high transition/transversion ratios, maximum likelihood was more accurate than neighbour-joining, and these methods were more efficient than parsimony methods.

While the Blosum62 and JTT models of substitution used with the ML method in PUZZLE are both suitable for amino acid data encoded by nuclear genes, the Blosum62 matrix is intended for use with more distantly related amino acid sequences (Strimmer and von Haeseler, 1999). The differences in the trees shown in figures 12 and 13 are probably due to the difference in the suitability of the model chosen to analyse that dataset. The relationships of plant RPB1 sequences are not as decisively resolved by the Blosum62 model as they are by the JTT model, so these sequences may be too closely related (or too conserved) for the Blosum62 model to work properly. In contrast, when the JTT and Blosum62 models are applied to a dataset of much more distantly related RPB1 D-F sequences, as shown in figures 17 and 18, only the Blosum62 model is able to relate slime molds as a sister group to green algae and plants. In this case, the JTT model could not resolve the relationships of any of the major groups of eukaryotes with one another.

Several other points of criticism arise with respect to the phylogenetic tree topologies reconstructed in this study. It was found previously (Wainright *et al*, 1993) that the absence of early branching lineages within each of the major groups to be classified, the use of partial sequences, and not using a closely and distinctively related outgroup (Rodrigo *et al*, 1994) all

detracted from the support for the grouping of animals and fungi as a monophyletic group and animals as monophyletic using 18S rRNA sequences as a phylogenetic marker. The phylogenetic trees presented here using regions D-F of RPB1 as a phylogenetic marker suffer from these problems.

Three criteria need to be improved to obtain a better estimate of the evolutionary relationships of land plants using RPB1. First of all, *Spirogyra* is not a good outgroup to the existing plant RPB1 D-F dataset because it is too distantly related to the ingroups. This is demonstrated in tables 5 and 6, which show that too much homoplasy exists between the outgroup and several ingroups. Second, this analysis lacks vascular plants such as *Marchantia, Equisetum* and *Psilotum*, which are earlier-diverging land plants and intermediate to the current outgroup and ingroups. Inclusion of representatives of other major lineages of gymnosperms and angiosperms, such as the Pinaceae, Rosidae and Asteridae would also help to resolve the relationships of the most basal angiosperm clades. Third, the use of short, partial sequences detracts from this analysis. Nei (1996) reiterates that "methods such as neighbor joining, likelihood, and parsimony methods produce reasonably good phylogenetic trees when a sufficiently large number of nucleotides or amino acids are used". Examination of the alignments in Appendices 1 and 2 shows few synapomorphies. This is particularly detrimental to parsimony analysis and resulted in multiple equally most parsimonious trees. A longer sequence from RPB1 would provide additional informative sites for analysis.

Since there is evidence that multiple copies of the RPB1 gene exist in *Magnolia* and *Nymphaea*, it is possible that this phylogenetic analysis suffers from the use of non-orthologous genes. These two plants are polyploids (6n and 12n), and it is unknown which copies of RPB1 identified by Southern analysis are expressed or which might be pseudogenes. The *Arabidopsis* and

Zea mays RPB1 sequences used here are from the expressed RPB1 gene, and *Arabidopsis* only has a single copy of RPB1 (Dietrich *et al*, 1990). All of the plant species used in this analysis are diploid except for *Magnolia* and *Nymphaea*. While the use of cDNA sequences of RPB1 for phylogenetic analysis would not exclude the possibility of comparing non-orthologous genes, it would be a good start and could be followed by a collection of more RPB1 sequence data from the same plants so that orthologous sequences might be found.

Several recent plant molecular evolution studies included a broad sampling of angiosperm taxa and several closely-related outgroups. They corroborated one another in finding that herbaceous dicots, including *Amborella* and *Nymphaea*, are the earliest extant angiosperm species. The parsimony analyses of nuclear-encoded phytochrome A and C genes from 26 angiosperms (Mathews and Dongohue, 1999) supported an herbaceous origin of angiosperms whether phytochrome A, C or A and C sequences were analysed. Parkinson *et al* (1999) performed a maximum-likelihood analysis, accounting for site-to-site rate heterogeneity, of a 6564 bp alignment of *rbcL* genes encoded in chloroplasts, small subunit rRNA genes encoded in the nucleus and small subunit rRNA, *coxI* and *rps2* genes encoded in mitochondria. This analysis included 51 taxa and had several gymnosperms (*Cycas*, *Ginkgo* and Coniferales) as the outgroups to angiosperms. Qiu *et al* (1999) also performed a parsimony analysis of a 8733 bp alignment of 18S rRNA genes encoded in the nucleus, *atp1* and *matR* genes encoded in mitochondria and *atpB* and *rbcL* genes encoded in chloroplasts of 105 angiosperm and gymnosperm species. Gnetales and several gymnosperms, including *Cycas*, *Ginkgo* and Coniferales were specified as the outgroup to angiosperms. Soltis *et al* (1999) performed an analysis similar to that of Qiu *et al* (1999), but less

the mitochondrial genes, which enabled them to compare sequences from 567 angiosperm and gymnosperm species.

While these most recent and large-scale studies provide very important information, it is clear that additional data from a protein-coding nuclear gene is required to corroborate these findings. Phylogenetic analysis of regions A through H of RPB1 would provide much longer sequences than the combined phytochrome dataset (Mathews and Donoghue, 1999) and would be informative when analysed independently as well as in combination with other nuclear, mitochondrial and chloroplast-encoded genes.

Substitution Analysis:

The substitution analysis of synonymous and non-synonymous sites of plant nucleotide sequences for regions D-F of RPB1 showed that the 278 synonymous sites had large levels of multiple substitutions, but the 556 non-synonymous sites had few if any (0-7) multiple substitutions. So, as long as synonymous sites are excluded, as they are here, little if any homoplasy is introduced into the analysis. The fact that there is very little multiple substitution at the nonsynonymous sites suggests that these sequences would be useful to resolve the evolutionary relationships of land plants if longer sequences were available.

Relative Rate Tests:

While the phylogenetic methods which account for substitution rate heterogeneity offer more resolution to the inference of the evolutionary relationships of plants with RPB1 D-F sequences, relative rate tests of this dataset do not indicate any significantly different rates of amino acid

substitution. Relative rate tests also show that regions D-F do not have significantly different substitution rates when comparing RPB1 of animals to plants to fungi, nor when comparing RPA1 to RPB1 to RPC1 in all eukaryotes.

Phylogenetic Analyses of Regions D-F of RPB1 in Other Eukaryotes:

Both the analyses of domains F-H and of domains D-F of RPB1 in animals, plants and fungi shown in figures 3 and in figures 14-16 consistently grouped animals and plants as sister groups, by parsimony, neighbor-joining, and PUZZLE maximum-likelihood methods. These results are contrary to most of the recent molecular phylogenetic studies of these groups, which support the evolutionary hypothesis that animals are more closely related to fungi than to plants.

Plants and animals were found to form a monophyletic group in molecular phylogenetic analyses of the 90 kDa heat shock protein Hsp90 (Gupta, 1995), multiple rRNA and tRNA species (Gouy and Li, 1989) and Cu-Zn superoxide dismutases (SOD) by Fitch and Ayala (1994). The analysis of Hsp90 was made by using neighbor-joining and parsimony methods to analyse a 620 amino acid alignment of 31 sequences, with approximately 40% identical sites. Gouy and Li's (1989) analysis was also made using neighbor-joining and parsimony methods. The analysis of SOD sequences by Fitch and Ayala (1994) using Fitch's ANCESTOR program and accounting for rate heterogeneity was made using an alignment of 67 amino acid sequences 118 residues in length, with the invariable sites removed. Fitch and Ayala point out that without examining the possibility that subsets of amino acids used in a phylogenetic analysis may evolve at different rates, errors can be made in estimating divergences by assuming a molecular clock. Gupta (1995) and Gouy and Li (1989) did not take this possibility into account while inferring those phylogenetic trees. A possible

problem with the SOD analysis, however, could be the shortage of informative sites in this relatively short and conserved sequence. The SOD dataset is even shorter in length than the RPB1 dataset presented in this study.

Sidow and Thomas (1994) also showed that animals and plants grouped together, excluding fungi, with distance, parsimony and maximum-likelihood analyses of RPB1 nucleotide and amino acid sequences. Alignments corresponding to the first and second bases of codons and the inferred amino acid sequence of 2038 codons of RPB1 were used for their analyses. These analyses were limited by a sparse sampling (6-10 species) of diverse eukaryotic taxa , however, and the authors did not report that they accounted for rate heterogeneity in their phylogenetic analysis.

Morris (1998) and references therein support some of the relationships seen amongst animals in figures 3 and 14, which are the maximum-likelihood tree of regions F-H of RPB1 and a parsimony tree of regions D-F of RPB1 in animals, plants and fungi. Morris' paper shows the consensus of fossil and molecular systematic evidence of the phylogenetic relationships of animals. The RPB1 F-H tree as well as the parsimony tree of the D-F data separate the deuterostomes (*Mus*) from the protostomes (the other animals). Figure 14 shows arthropods and nematodes grouping together, as a sister group to molluscs, which is also supported by Morris' paper, which also states that these lineages diverged more than 500 million years ago. Since this predates the origin of land plants, this is additional support to the argument that a longer sequence from RPB1 is needed to provide additional informative sites from this conserved protein sequence to sort out the evolutionary relationships of plants, since the D-F region does not even provide enough information to resolve the (earlier) relationships of animals with consensus using different methods.

Recent phylogenetic analyses of small subunit ribosomal RNA (SSU rRNA) sequences by maximum-likelihood methods (Wainright *et al*, 1993 and Van de Peer and De Wachter, 1997) and distance methods, accounting for rate heterogeneity amongst sites (Kumar and Rzhetsky, 1996), in addition to studies of numerous protein-coding genes all support a close relationship of animals with fungi, excluding plants from this group. Baldauf and Palmer (1993) found animals and fungi to be most closely related by both parsimony and distance analyses of sequences for EF-1α (an elongation factor), α and β–tubulin and actin. This relationship was also found in later analyses of these sequences, i.e. with neighbor-joining analysis of actin by Drouin *et al* (1995), with neighbor-joining analysis of α and β–tubulin by Keeling and Doolittle (1996) and with EF-1α using distance, parsimony and maximum-likelihood methods by Baldauf and Doolittle (1997) and Roger *et al* (1999). The latter two studies utilised the JTT amino acid transition probability matrix adjusted for amino acid frequences and the PROTML program for the maximum-likelihood analysis but did not account for rate heterogeneity. While these proteins also have highly conserved sequences, greater length (approximately 100-150 amino acids longer than this RPB1 D-F alignment) provided additional informative sites. Analysis of other elongation factors (EF-2 by Hirt *et al* (1999) and EF-Tu and EF-G by Baldauf *et al* (1996)) also supported the grouping of animals with fungi. Both analyses used distance, parsimony and PROTML maximum-likelihood methods, and account for rate heterogeneity within sequences, while Hirt *et al* (1999) also accounted for base composition heterogeneity between sequences in their analyses.

Sidow and Thomas (1994) point out that genes encoding elongation factors may undergo concerted evolution, so the evolutionary relationships of sequences of unrelated proteins should also be studied. Animals and fungi were also found to be sister groups in analyses of genes encoding the

60 kDa chaperonin (Cpn60) protein (Clark and Roger (1995), Roger *et al* (1998)), the 70 kDa heat-shock protein, Hsp70 (Borchiellini *et al* (1998)), aminoacyl-tRNA synthetases (Brown and Doolittle, 1995) and translation initiation factor eIF-2γ (Keeling *et al* (1998)).

These are only a few of some major examinations of the relationships amongst plants, fungi, and animals using molecular data. It is difficult to group two of these groups together with consensus because they have emerged during very proximal periods of time. This is also reflected by the weak support for many internal nodes in the analyses of RPB1 shown in figures 14-18.

The analyses represented by figures 14-18 were dominated by taxa which were very closely related amongst their own groups, and each of these groups were very distantly related to one another. This makes it likely that there is a shortage of synapomorphies amongst closely related taxa, while homoplasies have accumulated between distantly related groups. The topology of the RPB1 tree might be different if a more broad representation of sequences were included, including more early-branching lineages in each of the major eukaryotic kingdoms that would effectively be intermediate taxa amongst the major groups. This would reduce the homoplasies seen from one taxon to the next and might thus reduce the overall substitution rate heterogeneity. Unequal rates of substitution amongst organisms might skew the dataset to misrepresent the positions of taxa in a phylogenetic tree (Baldauf and Palmer, 1993).

The close relationship of RPB1 to RPA1 rather than to RPC1 shown in figures 17 and 18 is supported by neighbor-joining and maximum likelihood analyses of type A RNA polymerases by a set of complete sequences by Iwabe *et al* (1991), as well as by the parsimony and "parsimonious low sum of squares" trees inferred by Klenk *et al* (1995) from a 1168-amino acid alignment. The distance analysis by Klenk *et al* (1995) placed RPC1 and RPA1 as sister groups, but the analyses

in that paper were limited by a number of partial sequences, a sparse sampling of taxa and by not accounting for rate heterogeneity.

Phylogenetic analyses of domains D-G, A-G and A-H of RPB1 from 12-16 distantly related eukaryotes by Stiller and Hall (1997, 1998), Stiller *et al* (1998) and Hirt *et al* (1999) showed trees with stronger statistical support than the analyses of shorter RPB1 sequences (D-F) from a more broad selection of taxa (38 species) shown here in Figures 17 and 18. However, depending upon the selection of taxa, the alignment, and the method used to construct the trees in these studies, the relationships of animals, plants and fungi to one another were not always resolved in these analyses either, although some relationships of deeply diverging protist and algal lineages were always resolved. Stiller and Hall (1998) and Stiller *et al* (1998) showed that with consensus trees from parsimony, distance and maximum likelihood analyses of regions A-H of RPB1, red algae were distinctly separated from basal eukaryotic lineages and formed a sister group to the "crown" eukaryotic taxa. In this "crown" group, however, the relationships of animals, plants, fungi and slime molds to one another were not resolved. The relationships amongst the 3-4 most basal lineages (*Giardia, Trypanosoma, Mastigamoeba,* and *Trichomonas*) also were not well-resolved in these trees. Stiller and Hall (1997) found that phylogenetic trees of regions D-G of RPB1 inferred by both parsimony and distance methods found fungi as the sister group to a clade consisting of a plant, a green alga and a slime mold. Their maximum-likelihood analysis of regions A-H (excluding the green alga), however, placed fungi as a sister group to animals. Also, while parsimony and distance methods showed congruent relationships of basal eukaryotes (in considering regions D-G), maximum likelihood (considering regions A-H) resolved the topology of these relationships differently.

In the phylogenetic analysis of 15 sequences of regions A-G of RPB1 using a combination of parsimony and maximum likelihood methods, Hirt *et al* (1999) showed well-resolved relationships amongst the basal eukaryotic taxa, and a strongly supported relationship of Microsporidia with fungi. However, the fungi + Microsporidia clade was only placed as a sister group to animals with <50% bootstrap support. They also found that removal of fast-evolving sites from the alignment increased the bootstrap support in parsimony analyses, which are more sensitive to unequal substitution rates. For maximum-likelihood analysis, though, Hirt *et al* (1999) found that strong bootstrap support for deep relationships (amongst basal eukaryotic lineages) depended upon the inclusion of invariable sites in the alignment and was greatly reduced without them. These authors also found that by using methods which account for base composition heterogeneity between sequences and for site-by-site substitution rate heterogeneity, phylogenetic analyses of sequences for regions A-G of RPB1 yielded more conclusive results than did similar analyses of EF-1α and EF-2 sequences, which were more strongly affected by heterogeneous base composition and substitution rates.

The work of Stiller and Hall (1997, 1998), Stiller *et al* (1998) and Hirt *et al* (1999) further demonstrates several key issues about the use of RPB1 for studying evolutionary relationships. The analyses of regions A-G and A-H gave similar results, with relationships that were better resolved than the analyses of regions D-F shown in figures 17 and 18. These papers also demonstrate that the use of phylogenetic methods which account for base composition heterogeneity between sequences and for substitution rate heterogeneity within sequences when comparing RPB1 sequences of distantly related organisms provides a more robust estimate of their evolutionary relationships than can otherwise be obtained. These key points suggest that the use of longer (A-H)

RPB1 sequences, analysed using the methods of Hirt *et al* (1999) would be more useful for inferring the evolutionary relationships of land plants.

Conclusions:

Although a single and congruent phylogeny of land plants was not be found here due to the narrow representation of taxa in this analysis and the truncated and conserved sequences, it is likely that the phylogenies presented here using the maximum likelihood methods (figures 12 and 13) are the most accurate. Analysis of longer RPB1 sequences from a more broad range of taxa, taking site-to-site as well as taxon-to-taxon substitution rate heterogeneity into account, will be needed to further establish the usefulness of RPB1 for studying the evolutionary relationships of land plants. This should include more vascular plants, additional angiosperm and gymnosperm taxa, and a more closely related outgroup or outgroups. Regions A-D of RPB1 should also be included in such an analysis, because this region is less conserved than the D-F region and should provide additional informative nonsynonymous substitutions. Although domains D-F of RPB1 are too short to provide enough information to resolve the evolutionary relationships of the major groups of land plants, the non-synonymous substitution rate of 2×10^{-10} nonysnonymous substitutions per site per year (Figure 23 and Tables 9-12) and paucity of multiple substitutions at non-synonymous sites (Table 7) indicates that RPB1 could be a good phylogenetic marker to resolve the evolutionary relationships of land plants.

REFERENCES:

Ahearn J, Bartolmei M, West M, Cisek L, Corden J. (1987) Cloning and sequence analysis of the mouse genomic locus encoding th largest subunit of RNA polymerase II. *J Biol Chem* 262: 10695-10705.

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

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403-410.

Altschul SF, Madden T, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* 25:3389-3402

An SS, Möpps B, Weber K, Bhattacharya D. (1999) The origin and evolution of green algal and plant actins. *Mol Biol Evol* 16: 275-285.

Azuma Y, Yamagishi M, Ueshima R, Ishihama A (1991) Cloning and sequence determination of the *Schizosaccharomyces pombe rpb*1 gene encoding the largest subunit of RNA polymerase II. *Nucl Acids Res* 19: 461-468.

Baldauf SL, Palmer JD. (1993) Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc Natl Acad Sci USA* 90: 11558-11562.

Baldauf SL, Palmer JD, Doolittle WF (1996) The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc Natl Acad Sci USA* 93: 7749-7754.

Baldauf SL, Doolittle WF. (1997) Origin and evolution of the slime molds (Mycetozoa). *Proc Natl Acad Sci USA* 94: 12007-12012.

Baldwin JG, Frisse LM, Vida JT, Eddleman CD, Thomas WK. (1997) An evolutionary framework for the study of developmental evolution in a set of nematodes related to *Caenorhabditis elegans*. *Mol Phyl Evol* 8:249-259.

Bateman RM, Crane PR, DiMichele WA, Kenrick PR, Rowe NP, Speck T, Stein WE. (1998) Early evolution of land plants: Phylogeny, physiology and ecology of the primary terrestrial radiation. *Annu Rev Ecol Syst* 29: 263-92.

Bennett M.D., Smith J.B. (1976) Nuclear DNA amounts in angiosperms. 274: 227-274

Bennett M.D., Smith J.B., Heslop-Harrison J.S. (1982) Nuclear DNA amounts in angiosperms. *Proc. Roy. Soc. Lond.* 216: 179-199.

Bird DM, Riddle DL (1989) Molecular cloning and sequencing of *ama*-I, the gene encoding the largest subunit of *Caenorhabditis elegans* RNA polymerase II. *Mol Cell Biol* 9: 4119-4130.

Borchiellini C, Boury-Esnault N, Vacelet J, Parco YL. (1998) Phylogenetic analysis of the Hsp70 sequences reveals the monophyly of Metazoa and the specific phylogenetic relationships between animals and fungi. *Mol Biol Evol* 15: 647-655.

Bremer K (1985) Summary of green plant phylogeny and classification. *Cladistics* 1: 369-385.

Brown J, Doolittle WF. (1995) Root of the universal tree of life based on ancient aminoacyl-tRNA synthestase gene duplications. *Proc Natl Acad Sci USA* 92: 2441-2445.

Chang S, Puryear J, Cairney J. (1993) A simple method for isolating RNA from pine trees. *Plant Mol Biol Rep* 11:113-116

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu Y-L, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michaels HJ, Kress WJ, Karol KG, Clark WD, Hedren M, Gaut BS, Jansen RK, Kim K-J, Wimpee CF, Smith JF, Furnier GF, Strauss SH, Xiang Q-Y, Plunkett GM, Soltis PS, Swensen SM, Williams SE, Gadek PA, Quinn CJ, Eguiarte LE, Golenberg E, Learn GH Jr., Graham SW, Barrett SCH, Dayanandan S, Albert VA. (1993) Phylogenetics of seed plants: An analysis of nucleic acid sequences from the plastid gene *rbcL*. *Ann Miss Botan 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.

Cho Y, Qiu Y-L, Kuhlman P, Palmer JD. (1998) Explosive invasion of plant mitochondria by a group I intron. *Proc Natl Acad Sci USA* 95: 14244-14249.

Clark CG, Roger AJ. (1995) Direct evidence for secondary loss of mitochondria in *Entamoeba histolytica*. *Proc Natl Acad Sci USA* 92: 6518-6521.

Crane PR, Friis EM, Pedersen KR. (1995) The origin and early diversification of angiosperms. *Nature* 374: 27-33.

Croan D, Ellis J. (1996) Phylogenetic relationships between *Leichmania, Vianna* and *Sauroleishmania* inferred from comparison of a variable domain within the RNA polymerase II largest subunit gene. *Mol Biochem Parasit* 79: 97-102.

Cronquist A. (1981) *An integrated system of classification of flowering plants.* Columbia University Press, New York. p. 51.

Darlington CD, Ammal EKJ. (1945) *Chromosome atlas of cultivated plants.* George Allen & Unwin, London. pp. 38-63.

Denton AL, McConaughy BL, Hall BD. (1998) Usefulness of RNA polymerase II coding sequences for estimation of green plant phylogeny. *Mol Biol Evol* 15: 1082-1085.

Dietrich MA, Prenger JP, Guilfoyle T. (1990) Analysis of the genes encoding the largest subunit of RNA polymerase II in *Arabidopsis* and soybean. *Plant Mol Biol* 15: 207-223.

Donoghue MJ. (1994) Progress and Prospects in reconstructing plant phylogeny. *Ann Missouri Bot. Gard.* 81: 405-418.

Donoghue MJ, Mathews S. (1998) Duplicate genes and the root of angiosperms, with an example using phytochrome sequences. *Mol Phyl Evol* 9:489-500.

Doyle JJ, Doyle JL. (1990) Isolation of plant DNA from fresh tissue. *Focus* 12:13-15.

Drouin G, Prat F, Ell MJ, Clarke PTG. (1999) Detecting and characterizing gene conversion events between multigene family members. *Mol Biol Evol* 16:1369-1390.

Drouin G, Moniz de Sá M, Zuker M. (1995) The *Giardia lamblia* actin gene and the phylogeny of eukaryotes. *J Mol Evol* 41:841-9.

Evers R, Hammer A, Kock J, Jess W, Borst P, Memet S, Cornelissen AW (1989) *Trypanosoma brucei* contains two RNA polymerase II largest subunit genes with an altered C-terminal domain. *Cell* 56(4):585-97.

Felsenstein J. (1993) PHYLIP v3.52c, University of Washington, Seattle.

Fitch WM, Ayala FJ. (1984) The superoxide dismutase molecular clock revisited. *Proc Natl Acad Sci USA* 91:6802-6807.

Foster PG, Hickey DA (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol* 48:284-90 .

Garber AT. (1993) *Biofeedback* 14: 148-9

Gehrig HH, Heute V, Kluge M (1998) Toward a better knowledge of the molecular evolution of phosphoenolpyruvate carboxylase by comparison of partial cDNA sequences. *J Mol Evol* 46: 107-114.

Goremykin V, Bobrova V, Pahnke J, Troitsky A, Antonov A, Martin W. (1996) Noncoding sequences from the slowly evolving chloroplast inverted repeat in addition to *rbc*L data do not support Gnetalean affinities of angiosperms. *Mol Biol Evol* 13: 383-396.
Gouy M, Li WH. (1989) Molecular phylogeny of the kingdoms animalia, plantae, and fungi. *Mol Biol Evol.* 6: 109-122.

Gupta R. (1995) Phylogenetic analysis of the 90kD heat shock family of protein sequences and an examination of the relationship among animals, plants, and fungi species. *Mol Biol Evol.* 12: 1063-1073.

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.

Hasebe M, Kofuji R, Ito M, Kato M, Iwatsuki K, Ueda K. (1992) Phylogeny of gymnosperms inferred from *rbcL* gene sequences. *Bot Mag Tokyo* 105: 673-679.

Hasebe M, Ito M, Kofuji R, Ueda K, Iwatsuki K. (1993) Phylogenetic relationships of ferns deduced from *rbcL* gene sequence. *J Mol Evol* 37: 476-482.

Hasegawa, M., Hashimoto, T. (1993) rRNA trees misleading? *Nature* 361:23.

Henikoff S, Henikoff JG. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89: 10915-10919.

Higgins DG. (1994) CLUSTAL V: Multiple alignment of DNA and protein sequences. *Methods Mol Biol* 25: 307-18.

Hirt RP, Logsdon JM Jr, Healy B, Dorey MW, Doolittle WF, Embley TM. (1999) Microsporidia are related to fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc Natl Acad Sci USA* 96:

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

Jess W, Hammer A, Cornelissen AW (1989) Complete sequence of the gene encoding the largest subunit of RNA polymerase I of *Trypanosoma brucei. FEBS Lett* 249:123-8.

Jokerst RS, Weeks JR, Zehring WA, Greenleaf AL. (1989) Analysis of the gene encoding the largest subunit of RNA polymerase II in *Drosophila. Mol Gen Genet* 215:266-275.

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *CABIOS* 8: 275-282.

Keeling PJ, Fast N, McFadden GI. (1998) Evolutionary relationship between translation initiation factor eIF-2γ and selenocysteine-specific elongation factor SELB: change of function in translation factors. *J Mol Evol* 47: 649-655.

Keeling PJ, Doolittle WF. (1996) Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. *Mol Biol Evol* 13: 1297-1305.

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

Klenk HP, Zillig W. (1994) DNA-dependant RNA polymerase subunit B as a tool for phylogenetic reconstructions: branching topology of the archaebacterial domain. *J Mol Evol.* 38:420-432.

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

Knackmuss S, Bautz EF, Petersen G. (1997) Identification of the gene coding for the largest subunit of RNA polymerase I (A) of *Drosophila melanogaster. Mol Gen Genet* 253:529-34.

Kock J, Evers R, Cornelissen AW. (1988) Structure and sequence of the gene for the largest subunit of trypanosomal RNA polymerase III. *Nucl Acids Res* 16: 8753-8772.

Kolukisaoglu HU, Marx S, Wiegmann C, Hanelt S, Schneider-Poetsch HAW. (1995) Divergence of the phytochrome family predates angiosperm evolution and suggests that *Selaginella* and *Equisetum* arose prior to *Psilotum. J Mol Evol* 41: 329-337.

Kuhner MK, Felsenstein J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol* 11: 459-468.

Kumar S, Rzhetsky A. (1996) Evolutionary relationships of eukaryotic kingdoms. *J Mol Evol* 42: 183-193.

Kuzoff RK, Sweere JA, Soltis DE, Soltis PS, Zimmer EA. (1998) The phylogenetic potential of entire 26S rDNA sequences in plants. *Mol Biol Evol* 15: 251-263.

Lam TY, Chan L, Yip P, Siu CH. (1992) The largest subunit of RNA polymerase II in *Dictyostelium*: conservation of the unique tail domain and gene expression. *Biochem Cell Biol* 70: 792-799. Carleton QP501.C3222

Lanzendorfer M, Palm P, Grampp B, Peattie DA, Zillig W. (1992) Nucleotide sequence of the gene encoding the largest subunit of the DNA-dependant RNA polymerase III of *Giardia lamblia. Nucl Acids Res* 20:1145.

Li WB, Bzik DJ, Gu HM, Tanaka M, Fox BA, Inselburg J (1989) An enlarged largest subunit of *Plasmodium falciparum* RNA polymerase II defines conserved and variable RNA polymerase domains. *Nucl Acids Res* 17: 9621-36.

Li W-H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol* 36:96-9.

Li W-H, Graur D. (1991) *Fundamentals of Molecular Evolution.* Chapter 4. Sinauer, Sunderland MA.

Lloyd AT, Sharp PM. (1992) CODONS: a microcomputer program for codon usage analysis. *J Hered* 83: 239-240.

Loconte H. (1996) "Comparison of alternative hypotheses for the origin of the angiosperms", Chapter 10 in *Flowering plant origin, evolution and phylogeny.* Taylor DW, Hickey LJ, eds. Chapman and Hall, NY. 267-286.

Malek O, Lattig K, Hiesel R, Brennicke A, Knoop V. (1996) RNA editing in bryophytes and a molecular phylogeny of land plants. *EMBO* 15: 1403-1411.

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

Meacham C. (1994) Phylogenetic relationships at the basal radiation of angiosperms: further study by probability of character compatibility. *Syst. Bot.*. 19: 506-522.

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.

Mishler BD, Churchill SP. (1985) Transition to a land flora: phylogenetic relationships of the green algae and bryophytes. *Cladistics* 1: 305-328.

Moniz de Sá M, Drouin G. (1998) Phylogeny and substitution rates of one fern and six gymnosperm actin genes. unpublished

Moniz de Sá M, Drouin G. (1996) Phylogeny and substitution rates of angiosperm actin genes. *Mol Biol Evol* 13: 1198-1212.

Morris SC. (1998) Metazoan phylogenies: falling into place or falling to pieces? A paleontological perspective. *Curr Opin Genet Dev* 8:662-667.

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.

Nei M. (1996) Phylogenetic analysis in molecular evolutionary genetics. *Annu Rev Genet* 30: 371-403.

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

Parkinson CL, Adams KL, Palmer JD. (1999) Multigene analyses identify the three earliest lineages of extant flowering plants. *Curr Biol* 9: 1485-1488.

Palenik B (1992) Polymerase evolution and organism evolution. *Curr Opin Genes Dev* 2:931-936.

Ponce MR, Micol JL. (1992) PCR amplification of long DNA fragments. *Nucl Acids Res* 20:623

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

Puhler G, Lottspeich F, Zillig W. (1989b) Organization and nucleotide sequence of the genes encoding the large subunits A, B and C of the DNA-dependent RNA polymerase of the archaebacterium *Sulfolobus acidocaldarius*. *Nucl Acids Res* 17: 4517-4535.

Purugganan MD. (1997) The MADS-box floral homeotic gene lineages predate the origin of seed plants: Phylogenetic and molecular clock estimates. *J Mol Evol* 45: 392-396.

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

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

Qiu Y-L, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW. (1999) The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* 402: 404-407.

Quon D, Delgadillo M, Johnson P. (1996) Transcription in the early diverging eukaryote *Trichomonas vaginalis*: an unusual RNA polymerase II and alpha-amanitin resistant transcription of protein-coding genes. *J Mol Evol* 43: 253-62.

Regier JC, Shultz JW. (1997) Molecular phylogeny of the major arthropod groups indicates polyphyly of Crustaceans and a new hypothesis for the origin of hexapods. *Mol Biol Evol* 14: 902-913.

Ridley M. (1996) *Evolution*. 2nd ed. Blackwell, Cambridge MA. 537-538.

Ritland K, Eckenwalder J. (1992) Chapter 17: Polymorphism, hybridization, and variable evolutionary rate in molecular phylogenies. in *Molecular Systematics of Plants*, P. Soltis, D. Soltis, J, Doyle, eds. Chapman and Hall, New York. p. 404-429.

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

Rodrigo AG, Bergquist PR, Bergquist PL. (1994) Inadequate support for an evolutionary link between the metazoa and the fungi. *Syst. Biol.* 43: 578-584.

Roger AJ, Svärd SG, Tovar J, Clark CG, Smith MW, Gillin FD, Sogin ML. (1998) A mitochondrial-like chaperonin 60 gene in *Giardia lamblia*: Evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria. *Proc Natl Acad Sci USA* 95: 229-234.

Roger AJ, Sandblom O, Doolittle WF, Philippe H. (1999) An evaluation of elongation factor 1 α as a phylogenetic marker for eukaryotes. *Mol Biol Evol* 16: 218-233.

Rose T, Schultz E, Henikoff J, Pietrokovski S, MacCallum C, Henikoff S. (1998) Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucl Acids Res* 26: 1628-1635

Sambrook J, Fritsch EF, Maniatis T. (1989) *Molecular cloning: a laboratory manual*, 2nd ed. Cold Spring Harbor Laboratory Press.

Sandhu GS, Precup JW, Kline BC. (1989) Rapid one-step characterization of recombinant vectors by direct analysis of transformed *Escherischia coli* colonies. *BioTechniques* 7: 689-690.

Seither P, Coy JF, Pouska A, Grummet I. (1997) Molecular cloning and characterization of the cDNA encoding the largest subunit of mouse RNA polymerase I. *Mol Gen Genet* 255: 180-186.

Sidow A, Thomas WK. (1994) A molecular evolutionary framework for eukaryotic model organsims. *Curr Biol.* 4: 596-603.

Sidow A, Wilson A. (1990) Compositional statistics: an improvement of evolutionary parsimony and its application to deep branches in the tree of life. *J Mol Evol* 31: 51-68.

Smith SW, Overbeek R, Woese CR, Gilbert W, Gillevet PM (1994) The genetic data environment: an expandable GUI for multiple sequence analysis. *Comput Appl Biosci* 10:671-5.

Soltis DE, Soltis PS, Nickrent DL, Johnson LA, Hahn WJ, Hoot SB, Sweere JA, Kuzoff RK, Kron KA, Chase MW, Swensen SM, Zimmer EA, Chaw S-M, Gillespie LJ, Kress WJ, Sytsma KJ. (1997) Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. *Ann Miss Bot Gard* 84: 1-49.

Soltis PS, Soltis DE, Chase MW. (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 402: 402-404.

Sparrow AH, Price HJ, Underbrink AG. (1972) A survey of DNA content per cell and per chromosome of prokaryotic and eukaryotic organisms: some evolutionary considerations. *Evolution of genetic systems*, ed by HH Smith. Gordon and Breach, NY. p.451-494.

Stebbins GL Jr. (1950) *Variation and Evolution in Plants*. Columbia University Press. NY. pp. 298-355.

Stewart WN, Rothwell GW. (1993) Chapter 32: Major evolutionary events and trends - in retrospect, in *Paleobotany and the evolution of plants*. Cambridge University Press, NY. pp. 505-512.

Stiller JW, Hall BD. (1997) The origin of red algae: implications for plastid evolution. *Proc Natl Acad Sci USA* 94:4520-4525.

Stiller JW, Duffield CS, Hall BD. (1998) Amitochondriate amoebae and the evolution of DNA-dependant RNA polymerase II. *Proc Natl Acad Sci USA* 95: 11769-11774.

Stiller JW, Hall BD. (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.

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

Strimmer K, von Haeseler A. (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci USA* 94: 6815-6819.

Strimmer K, von Haeseler A. (1999) Manual for PUZZLE 4.0.2, distributed with PUZZLE as: puzzle4.0.2\docs\manual.html

Takhtadzhian AL. (1997) Introduction: Evolutionary Systematics: The Darwinian paradigm, in *Diversity and classification of flowering plants*. Columbia University Press, NY. pp. 1-7.

Tateno Y, Takezaki N, Nei M. (1994) Relative efficiencies of the maximum likelihood, neighbor-joining, and maximum parsimony methods when substitution rate varies with site. *Mol Biol Evol* 11: 261-277.

Taylor DW, Hickey LJ (1996) Chapter 1, "Introduction: The challenge of flowering plant history" and Chapter 9, "Evidence for and implications of an herbaceous origin for angiosperms" in *Flowering plant origin, evolution and phylogeny*. Taylor DW, Hickey LJ, eds. Chapman and Hall, NY. 1-7, 232-266.

Thompson JD, Higgins DG, Gibson TJ. (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.

Van de Peer Y, De Wachter R. (1997) Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA. *J Mol Evol* 45: 619-630.

Wainright PO, Hinkle G, Sogin ML, Stickel SK. (1993) Monphyletic origins of the metazoa: an evolutionary link with fungi. *Science* 260: 340-342.

Winter K-U, Becker A, Münster T, Kim JT, Saedler H, Theissen 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.

Wintzerith M, Acker J, Vicaire S, Vigneron M, Kedinger C. (1992) Complete sequence of the human RNA polymerase II largest subunit. EMBL-Genbank database, Accession number X63564. *Nucl Acids Res.* 20:910.

Yamagishi M, Nomura M. (1988) Cloning and sequence determination of the gene encoding the largest subunit of the fission yeast *Schizosaccharomyces pombe* RNA polymerase I. *Gene* 74:503-515.

Young RA (1991) RNA polymerase II. *Annu Rev Biochem* 60: 689-715.

Appendix 1: 556 bp alignment of 9 nucleotide sequences corresponding to the first and second bases of codons in the region of the RPB1 gene of plants which encodes domains D-F, which was used to infer the phylogenetic trees shown in figures 6-8. "N" represents positions where "Y" and "m" were substituted for the first base of leucine and arginine codons. *Spirogyra* is the outgroup. Residues identical to the outgroup sequence are indicated by a dot.

```
                                                                                                          100
Spirogy.II   CAGTTGCAACTTGAACnGGC GAACATGAnTATATGTCCAA TGGTGTAGCCCATCAAnGCC GTATGGATGTCAGAACnTnT GGTGnGAAGTACAAnGGAAC
Arau.hetII   ....CC..T..........G ..GTGC......n....... ..A...TC............ .................... ....................
Cyc.rev.II   ....CC..T........... ..GTn............... ..A...TC............ .................... ........A...........
Eph.vir.II   ....CC..T........... ...Tn............... ..A...TC............ .................... ........A...........
Gnk.bil.II   ....CC..T........... ..GTn............... ..A...TC..A......... .................... ....................
Mag.sou.II   ....CC..T........... ..GTn............... ..A...TC............ .................... ........A...........
Nym.odo.II   ....CC..TG.......... ..GTn............... ..A...TC............ .................... ........A...........
Zea.mays.2   ....CC..T........... ..GTn............... ..A...TC............ .................... ........A...........
Arab.tha.2   ....CC..T........... ..GTn............... ..A...TC...G........ .................... ........A...........

                                                                                                          200
Spirogy.II   TTATGAAAGAGTTTATAAAT nTATTGTGGAGATTGAGGAA ATCCTCCCACATnTAACCnG CCnTTGACGGAACAGTTTTC nTATATCCnGGCGTAAnTGA
Arau.hetII   .................... .................... ....A..G............ ..A...........A...AG ........AACAA...A.AC
Cyc.rev.II   .................... .................... .................... ..................AA ..........CAA.....AT
Eph.vir.II   .................... .................... G...G...G........... ..A...............AA ........AACAA......T
Gnk.bil.II   .................... .................... .................... ..................AA ..........CAA.....AT
Mag.sou.II   .................... .................... ....G............... ..................AA ........AACAA.....AT
Nym.odo.II   .................... .................... ....G...G...A....... ..................AA ........AACAA.....AT
Zea.mays.2   n................... ........C........... ....G............... ..A...............AA ........AACAA.....AT
Arab.tha.2   ...................C .................... G...G...G........... ..................AA ........AACAA.....nT

                                                                                                          300
Spirogy.II   nGTATCGCTGCACCGATCGA nGGGGATTTCCCGGGAACCA GTnGGTGAAAGGGAnTnTGC GGATnTTGAAAATCnTGGAC TCGGGGAGnTGTCAATATTG
Arau.hetII   ..ACG.......AA..AG.. A.AC..G.A.........AG ....A.............A. ...C.......A........ ..........A...G.....
Cyc.rev.II   ...........T........ AC..TTA.A.........TG ....A.........G...T. ...C.......A........ ..TC......A...G.....
Eph.vir.II   ...........AA..AG... ....C.A.A.nT......GT ....A.........G.A.A. ...C.......A....G... ..A.......A...G.....
Gnk.bil.II   ...........T........ AC..TTA.A.........TG ....A.........G...T. ...C.......A........ ..TC......A...G.....
Mag.sou.II   ..AC.......T...G.... AC..TTA.A........... ....A...nG.......... ...C.......A........ ..AC......A...G.....
Nym.odo.II   ...........T........ AC..TTA.A.........TG ....A...nG.........T. ...C.......A.A...... ..TC......A...G.....
Zea.mays.2   ...T.......T...GA... AA..TTA.A.........AT ....A.............T. ...C.......AG....... GGTC......A...G.....
Arab.tha.2   ...........G...A.... AC..TTA.A........... ....A...nG.......... ...C.......A........ ..AA..........G.....
```

```
                                                                                                  400
Spirogy.II   GAGAGTGGCCGAGCGCnGAA TTnTGGCAACCATGnTGTAA TATGnTnTCACAGGTTAGAT GGATGGGAACATGCGAGCTC ACATGAACATAAGAACATGC
Arau.hetII   .................... .................... .................... ................T.G. .......A............
Cyc.rev.II   .................... .................... .................... ..................G. ..................T.
Eph.vir.II   .................... .................... .................... ..................A. ..................CA
Gnk.bil.II   .................... .................... .................... ..................G. ..................T.
Mag.sou.II   .................... .................... ..........A......... .................... .......A..........T.
Nym.odo.II   .................... .................... ..........A..C...... .................... .......A..........T.
Zea.mays.2   .................... .................... ..........A......... .................... ..................T.
Arab.tha.2   .................... .................... ..........A......C.. ................T... .......A..........T.

                                                                                                  500
Spirogy.II   AAGCAAACGAGTAAGAnTAT GAGCGCTGGAAACAnTGAGC CACCGGnGACnTATGATCTT GAAAnGGTAACAGTnTAAAA GCnGGAGAGCGGnGGCGCCA
Arau.hetII   ......CA............ A...T.CA....AG....C. .................... ....AA............nG ............A.AG....
Cyc.rev.II   ......G.......C..... C.nT..CAC...GC...... G.........A......... .................... ............A.AG....
Eph.vir.II   G......TA...C....... ..AATAAT.CC.A.....CA G...........n....... ....CA.............. ............AAAG....
Gnk.bil.II   .......A......C..... A.....CA....GC...... G.........A......... .................... ............A.AG....
Mag.sou.II   .......A............ A.....CA............ G.........A......... .................... ............A.AG....
Nym.odo.II   .......A............ A.....CA............ G.........A......... .................... ............A.AG....
Zea.mays.2   .......A.C.......... A.AA..CA............ G.........A......... .................... ............A.AG....
Arab.tha.2   .........C.......... nGCATTCA.G..G.....C. G.........A.nG..A... .................... ............A.AG....

                                                          556
Spirogy.II   TCAGnTTCGAAGAAAAGTAA GCATGTACGCGGTCAAGGTC TTATAAATTCCAATAT
Arau.hetII   AA.................. .................... ................
Cyc.rev.II   AA..............n... ..................AG ...............C
Eph.vir.II   nG..............n... ..................AG ...............C
Gnk.bil.II   nG..............n... ..................AG ...............C
Mag.sou.II   AA..............n... ..................AG ...............C
Nym.odo.II   nG..............n... ..................AG ...............C
Zea.mays.2   AA..............n... ..................AG ...............C
Arab.tha.2   AA....G....C....n... ..................AG ...............C
```

Appendix 2: 278 position alignment of 9 amino acid sequences of regions D-F of RPB1 in land plants that was used to infer the phylogenetic trees shown in figures 10-13. A green alga, *Spirogyra*, is the outgroup. Residues identical to those in the outgroup are represented by a dot.

```
                                                                                                          100
Spirogyra  HVCQTFETRAETMELMMVPK CVVSPQSNRPVMGIVQDTLL GCRKVTKRDTFIEKDVFMNI LMWWEDFEGKIPSPTILKPR PLWTGKQVFSLIIPRAVNLE
Arau.het   ..P.S....G.VA...L... .I.........I........ .................... ....D..D..M.T.A..... .I.....I......KQI.MT
Cyc.rev    ..P.S......VL....... .I.................. ....I............... .......D............ .........N.....QI..I
Eph.vir    ..P.S......IL....... .I.................. ....I............... .......D..V.A.A..... .I.......N....KQI..V
Gnk.bil    ..P.S......VL....... .I...K.............. .................... .......D............ .........N.....QI..I
Mag.sou    ..P.S......VL....... .I.................. ....I............... .......D....A....... .........N....KQI..I
Nym.odo    ..P.C......VL....... .I.................. ....I............... .......D....A.A.M... .........N....KQI..I
Zea mays   ..P.S......VL....... .I.................. ....I.....L......... ....Q..D....A....... .I.......N....KQI..I
Arab.tha   ..P.S......VL....... .I....A............. ....I..............T .......D..V.A.A..... .........N....KQI..L

                                                                                                          200
Spirogyra  RYSAWHPDSERGDFSPGDTQ VRVEKGELLAGILCKKSLGT SGGSLVHIIWEEVGPDAARK FLGHTQWLVNYWLLQQGFSI GIGDTIADASTMDTINETIA
Arau.het   .TA...N...ST.VT....S ..I......T.T....T... .....I.V............ ...............H.... ........SA..EK......
Cyc.rev    ......SE..T.FIT....C ..I....V.S.T....T... .S...I.V............ .................... .........A..E......S
Eph.vir    ......NE.D..HMTL...V ..I....VIT.T....T..A .S...I.V............ ...................M .........T.........Q
Gnk.bil    ......SE..T.FIT....C ..I....V.S.T....T... .S...I.V............ .................... .........A..E......S
Mag.sou    .T....SEA.T.FIT..... ..I.R......T....T... .T...I.V............ ...............N.... ............EK.....S
Nym.odo    ......SE..T.FIT....C ..I.R....S.T....TM.. .S...I.V............ ...............NA... ............EK.....S
Zea mays   .F....SEE.K.FIT....M ..I......S.T........ GS...I.V............ ...............N.... ............E...D..S
Arab.tha   ......A.T.T.FIT..... ..I.R......T....T... .N....V............. ...............N..T. ........S...EK.....S

                                                                                      278
Spirogyra  KAKTEVKDLIEAACEKQLEA QPGRTLMESFENRVNQVLNK ARDDAGRAAQSSLSESNNVK AMVTAGSKGSFINISQMI
Arau.het   ...QD..E..K.SQ..S..P ............K......R ...E..SS..K......... ..................
Cyc.rev    ...A..NQ..QL.HQ.A... E....M.............. ......SS..K.......L. .................T
Eph.vir    D..IK.QE...KYMAHK..Q E.....L.....Q....... ......NS..R.......L. .................T
Gnk.bil    ...N...Q..K..Q..A... E....M.............. ......SS..R.......L. .................T
Mag.sou    ...N...E..K..Q...... E....M.............. ......SS..K.......L. .................T
Nym.odo    ...N...E..K..Q...... E....M.............. ......SS..R.......L. .................T
Zea mays   ...NA..E..KK.H...... E....M.............. ......SS..N.......L. .................T
Arab.tha   N...A.....RQFQG.E.DP E....MRDT........... ......SS..K..A.T..L. .................T
```

Appendix 3: 282 position alignment of 19 amino acid sequences of regions D-F of RPB1 in animals, plants and fungi that was used to infer the phylogenetic trees shown in figures 14-16. A red alga, *Porphyra yezeonsis*, is the outgroup. Residues identical to those in the outgroup are represented by dots while gaps in the alignment are shown as dashes.

```
                                                                                                               100
Por.yez     HVPQTHATRAEVMELMMVPR CIVSPQGNKPVMGIVQDTLL GCMLFTYRDTFLRRDVTMSL LLHVEGWDGVIPPPAIIKPE PLWTGKQLFSLLLP-DVNLV
Sac.cer     ....SEE....LSQ.CA..L Q.....S...C........C .IRKL.L....IEL.QVLNM .YW.PD......T......K ...S...IL.VAI.NGIH.Q
Sch.pom     ....SEE....IQ.IT...K Q.....S............A .VRK.SL..N..T.NAV.NI M.W.PD...IL...V.L..K V......IL..II.KGI..I
Mag.sou     ....SFE.....L......K ......S.R........... ..RKI.K....IEK..F.NI .MWW.DF..K..A.T.L..R .......V.N.II.KQI..I
Nym.odo     ....CFE.....L......K ......S.R........... ..RKI.K....IEK..F.NI .MWW.DF..K..A...M..R .......V.N.II.KQI..I
Zea mays    ....SFE.....L......K ......S.R........... ..RKI.K...LIEK..F.NI .MWWQDF..K..A.T.L..R .I.....V.N.II.KQI..I
Cyc.rev     ....SFE.....L......K ......S.R........... ..RKI.K....IEK..F.NI .MWW.DF..K..S.T.L..R .......V.N.II.RQI..I
Gnk.bil     ....SFE.....L......K .....KS.R........... ..RKV.K....IEK..F.NI .MWW.DF..K..S.T.L..R .......V.N.II.RQI..I
Arab.tha    ....SFE.....L......K ......A.R........... ..RKI.K....IEK..F.NT .MWW.DF..KV.A...L..R .......V.N.II.KQI..L
Eph.vir     ....SFE....IL......K ......S.R........... ..RKI.K....IEK..F.NI .MWW.DF..KV.A...L..R .I.....V.N.II.KQI...
Spirogyra   ..C..FE....T.......K .V....S.R........... ..RKV.K....IEK..F.NI .MWW.DFE.K..S.T.L..R .......V...II.RA...E
Arau.het    ....SFE..G..A...L..K ......S.R..I........ ..RKV.K....IEK..F.NI .MWWDDF..KM.T...L..R .I.....I...II.KQI.MT
Dro.mel     ....SME.....ENIHIT.. Q.IT..A............T AVRKM.K..V.IT.EQV.N. .MFLPT..AKM.Q.C.L..R .......I...II.GN..MI
Art.sal     ..A.SLE....LENIHIT.. Q.IT..A.R..........C AVRKM.K..V..EKEQM.T. .MYLPT...KL.Q...L..K .........T.II.GK..MM
Cra.gig     .L..SLE.K..ISN.AL... M.IT..A.R..........T AVRKM.K..V..D.GQX.N. .MFLPR...HV.Q...L..K ..........VI.GRT.CI
Ily.obs     .L..SLE....I.N.CA... M.IT..A.R..........T AVRKM.K..V..T.AQM.H. .MFLPT...RM.Q....... .......I...VI.GR..VI
Mus.mus     .L..SLE....IQ..A.... M..T..S.R..........T AVRK..K..V..E.GEV.N. .MFLST...KV.Q...L..R .......I...II.GHI.CI
Cae.ele     .L..SLE....IE.IA.... QLIT..A............C AVRMM.K..V.IDWPFM.D. .MYLPT...KV.Q...L..K .......V...II.GN..VL
Hel.sta     .LA.SLE....ISQ.AS.K. M.IT..A.R........S.T AVNKM.R....ITK.EI.NI .MYLP..A.KL.Q...L..R A.........III.GRI.VI
```

```
                                                                                                                          200
Por.yez     RYCNTHPDDE--S--TDISP GDTRVLIVGGELITGIVDKR TVGSAANGLIHVTWKEKGPE RTCVLISAIQVLVNHYVIMR GQSIGIGDTIADAHTDANVR
Sac.cer     .FDEGTTLLS---------. K.NGM..ID.QI.F.V.E.K ....SNG.....VTR....Q VCAK.FGN..KV..FWLLHN .F.T........GP.MREIT
Sch.pom     .DDDKQSLSN---------. T.SGM..EN..I.Y.V...K ...ASQG..V.TI....... ICKGFFNG..RV..YWLLHN .F...........D.MKE.T
Mag.sou     .TSAW.SEA.-TG---F.T. ...Q.R.ER...LA.TLC.K .L.TSTGS....I.E.V..D AARKFLGHT.W...YWLLQN .F...........S.MEKIN
Nym.odo     ..SAW.SES.-TG---F.T. ...C.R.ER...LS.TLC.K .M.TSSGS....I.E.V..D AARKFLGHT.W...YWLLQN AF...........S.MEKIN
Zea mays    .FSAW.SEE.-KG---F.T. ...M.R.EK...LS.TLC.K SL.TGSGS....I.E.V..D AARKFLGHT.W...YWLLQN .F...........S.METIN
Cyc.rev     ..SAW.SES.-TG---F.T. ...C.R.EK..VLS.TLC.K .L.TSSGS....I.E.V..D AARKFLGHT.W...YWLLQQ .F...........A.METIN
Gnk.bil     ..SAW.SES.-TG---F.T. ...C.R.EK..VLS.TLC.K .L.TSSGS....I.E.V..D AARKFLGHT.W...YWLLQQ .F...........A.METIN
Arab.tha    ..SAW.A.T.-TG---F.T. ...Q.R.ER...LA.TLC.K .L.TSNGS.V..I.E.V..D AARKFLGHT.W...YWLLQN .FT.........SS.MEKIN
Eph.vir     ..SAW.NESD-RG---HMTL ...V.R.EK..V...TLC.K .L.ASSGS....I.E.V..D AARKFLGHT.W...YWLLQQ .F.M.........T.MDTIN
Spirogyra   ..SAW...S.-RG---.F.. ...Q.RVEK...LA..LC.K SL.TSGGS.V.II.E.V..D AARKFLGHT.W...YWLLQQ .F...........S.MDTIN
Arau.het    .TAAW.N.S.-ST---.VT. ...S.R.EK...L..TLC.K .L.TSGGS....I.E.V..D AARKFLGHT.W...YWLLQH .F..........SA.MEKIN
Dro.mel     .THS....E.DEGPYKW... ...K.MVEH....M..LC.K SL.TS.GS.L.ICFL.L.HD IAGRFYGN..TVI.NWLLFE .H..........PQ.YNEIQ
Art.sal     KTHS......DDGPYKW... ...K.MVEH....M..LC.K .L.AX.GSIL.IIFL.L.HD ICGKFXGN..TV..NWLLYE .H..........PQ.YNSIQ
Cra.gig     .THS....E.DKGPYKW... ...K...ED.M..S..LC.K .L.TSSGS.A..VFM.Y.WV IAGEMYGH..T...NWLLLE .H..........PQ.YIDIQ
Ily.obs     .THS....G.DSGPYKW... ...K...ED....S..LC.K .L.TS.GS.V.IVFL.M.F. VAGE.YGN..TV..NWLLIE .H..........QQ.YHEIQ
Mus.mus     .THS......DSGPYKH... ...K.VVEN....M..LC.K SL.TS.GS.V.ISYL.M.HD I.RLFY.N..TVI.NWLLIE .HT.....S...SK.YQDIQ
Cae.ele     .THS....S.DSGPYKW... ...K.I.EH...LS...CSK ...KS.GN.L..VTL.L.Y. IAANFY.H..TVI.AWL.RE .HT.........QA.YLDIQ
Hel.sta     .THS......DRGPHKW... ...K..VED.K.LS..LC.K SL.ASSGS.Q.IIHH.L.SD A.ADFYAY..MVT..WLLVT .HT...A......K.YSDIQ

                                                                                      282
Por.yez     ATITGAQAEVTQLERRAQEG ELTLLPGKSMMESFEVEVNK VLNGARDTSGSSAQLSLLKS NNIKRMVSAGSKGSFINISQ IC
Sac.cer     E..AE.KKK.LDVTKE..AN L..AKH.MTLR....DN.VR F..E...KA.RL.EVN.KDL ..V.Q..M..........A. MS
Sch.pom     R.VKE.RRQ.AECIQD..HN R.KPE..MTLR....AK.SR I..Q...NA.R..EH..KD. ..V.Q..A............ MS
Mag.sou     E..SK.KN..KE.IKA...K Q.EAE..RT......NR..Q ...K...DA.....K..SE. ..L.A..T............ MT
Nym.odo     E..SK.KN..KE.IKA...K Q.EAE..RT......NR..Q ...K...DA.....R..SE. ..L.A..T............ MT
Zea mays    D..SK.KNA.KE.IKK.H.K Q.EAE..RT......NR..Q ...K...DA.....N..SE. ..L.A..T............ MT
Cyc.rev     E..SK.K...N..IQL.HQK A.EAE..RT......NR..Q ...K...DA.....K..SE. ..L.A..T............ MT
Gnk.bil     E..SK.KN..K..IKA...K A.EAE..RT......NR..Q ...K...DA.....R..SE. ..L.A..T............ MT
Arab.tha    E..SN.KTA.KD.I.QF.GK ..DPE..RT.RDT..NR..Q ...K...DA.....K..AET ..L.A..T............ MT
Eph.vir     E..QD.KIK.QE.IEKYMAH K.EQE..RTLL....NQ..Q ...K...DA.N...R..SE. ..L.A..T............ MT
Spirogyra   E..AK.KT..KD.IEA.C.K Q.EAQ..RTL.....NR..Q ...K...DA.RA..S..SE. ..V.A..T............ MI
Arau.het    E..AK.KQD.KE.IKAS..K S.EPQ..RTL.....NK..Q ...R...EA.....K..SE. ..V.A..T............ MI
Dro.mel     QA.KK.KDD.INVIQK.HNM ..EPT..NTLRQT..NK..R I..D.H.KT.G..KK..TEY ..L.A..VS.....N..... VI
Art.sal     T..KK.KED.IDVITK.HNN ..EPT..NTLRQT..NQ..R I..D...KT.G..KN..TEY ..L.A..VS.....N..... VI
Cra.gig     D..KK.KHD.IEVIEK.HND D.EPT..NTLRQT..NM..R I..D...KT..K..K..SDY ..F.A..V......K..... VI
Ily.obs     E..RK.K.D.IEVIEK.HND ..EPT..NTLRQT..NQ..R I..D...KT..K..K..SEF ..F.A..V......K..... VI
Mus.mus     N..KK.KQD.IEVIEK.HNN ..EPT..NTLRQT..NQ..R I..D...KT.....K..SEY ..F.S..VS.A...K..... VI
Cae.ele     N..RK.KQD.VDVIEK.HND D.EPT..NTLRQT..NK..Q I..D...RT.....K..SEF ..F.S..VS.....K..... VI
Hel.sta     TA.KK.KSD.VEVIEK.HND ..EPM..NTLRQT..NQ..R I..D...KT..L..K..SEF ..F.S..V.....NK..... VI
```

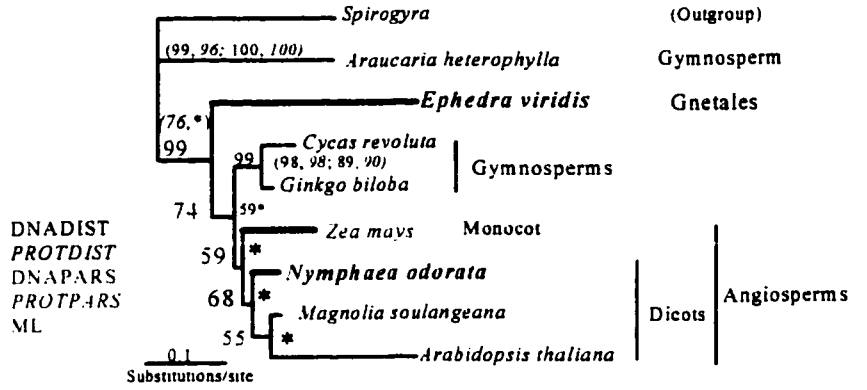
Appendix 4: 252 position alignment of 38 amino acid sequences of regions D-F of RPA1, RPB1 and RPC1 from diverse eukaryotes that was used to infer the phylogenetic trees shown in figures 17 and 18. *Sulfolobus acidocaldarius*, a member of the Archaea, is the outgroup. Residues identical to those in the outgroup are represented by dots while gaps are shown as dashes.

```
                                                                                                         100
Sulf.acido     NLHVPQSEEAIAEARELMLV HKNIITPRYGGPIIGGGQDY ISGAYLLSVKTTLLTVEEVA TILGVTELGEPAILAPKPYY TGKQVISLFLPDSFIVIKNG
Por.yez.II     ......THATR..VM...M. PRC.VS.QGNK.VM.IV..T LL.CM.FTYRD.F.RRDVTM SL.LHVVIPP...IK.E.LW ....LF..L...TRVL.VG.
Bon.ham.II     ......THQTR..VQ..... PHC.VS.QGNK.VM.IV..T LL.CM.FTQRD.F.ERDLMM NL.MHVVIPK...YK.R.LW .....F..I...TKVY.SR.
Dro.mel.II     .......M.TR..VENIHIT PRQ....QANK.VM.IV..T LTAVRKMTKRDVFI.R.Q.M NL.MFLKMPQ.C..K.R.LW ....IF..II..TKVMVEH.
Art.sal.II     ....A..L.TR..LENIHIT PRQ....QANR.VM.IV..T LCAVRKMTKRDVF.EK.QMM .L.MYLK.PQ....K...LW ....LFT.II..TKVMVEH.
Cra.gig.II     ...L...L.TK..ISN.A.. PRM....QANR.VM.IV..T LTAVRKMTKRDVF.DRGQXM NL.MFLHVPQ....K...LW ....LF..VI..TKVL.ED.
Ily.obs.II     ...L...L.TR..IMN.CA. PRM....QANR.VM.IV..T LTAVRKMTKRDVF..RAQMM HL.MFLRMPQ...IK.E.LW ....IF..VI..TKVL.ED.
Mus.mus.II     ...L...L.TR..IQ..AM. PRM.V..QSNR.VM.IV..T LTAVRKFTKRDVF.ERG..M NL.MFLKVPQ....K.R.LW ....IF..II..TKV.VE..
Cae.ele.II     ...L...L.TR..IE.IAM. PRQL...QANK.VM.IV..T LCAVRMMTKRDVFIDWPFMM DL.MYLKVPQ....K...LW .....F..II..TKVI.EH.
Hel.sta.II     ...LA..L.TR..ISQ.AS. KRM....QANR.VM.IV..S LTAVNKMTRRD.FI.KD.IM N..MYLK.PQ....K.RALW ....LF.III..TKVLVED.
Spirogyra.II   .M..C.TF.TR..TM...M. P.CVVS.QSNR.VM.IV..T LL.CRKVTKRD.FIEKDVFM N..MWWKIPS.T..K.R.LW .....F..II..TQVRVEK.
Arau.het.II    .M.....F.TRG.VA..... P.C.VS.QSNR.V..IV..T LL.CRKVTKRD.FIEKDVFM N..MWWKMPT....K.R.IW ....IF..II..TSVR.EK.
Cyc.rev.II     .M.....F.TR..VL...M. P.C.VS.QSNR.VM.IV..T LL.CRKITKRD.FIEKDVFM N..MWWKIPS.T..K.R.LW .....FN.II..TCVR.EK.
Eph.vir.II     .M.....F.TR..IL...M. P.C.VS.QSNR.VM.IV..T LL.CRKITKRD.FIEKDVFM N..MWWKVPA....K.R.IW .....FN.II..TVVR.EK.
Gnk.bil.II     .M.....F.TR..VL...M. P.C.VS.KSNR.VM.IV..T LL.CRKVTKRD.FIEKDVFM N..MWWKIPS.T..K.R.LW .....FN.II..TCVR.EK.
Mag.sou.II     .M.....F.TR..VL...M. P.C.VS.QSNR.VM.IV..T LL.CRKITKRD.FIEKDVFM N..MWWKIPA.T..K.R.LW .....FN.II..TQVR.ER.
Nym.odo.II     .M....CF.TR..VL...M. P.C.VS.QSNR.VM.IV..T LL.CRKITKRD.FIEKDVFM N..MWWKIPA...MK.R.LW .....FN.II..TCVR.ER.
Zea.mays.2     .M.....F.TR..VL...M. P.C.VS.QSNR.VM.IV..T LL.CRKITKRD..IEKDVFM N..MWWKIPA.T..K.R.IW .....FN.II..TMVR.EK.
Arab.tha.2     .M.....F.TR..VL...M. P.C.VS.QANR.VM.IV..T LL.CRKITKRD.FIEKDVFM NT.MWWKVPA....K.R.LW .....FN.II..TQVR.ER.
Dic.dis.II     ......TL.TR..VI.I.M. PRQ.VS.QSNR.VM.IV..T LL.SR.FTKRDCFMEKDL.M N..MWLKVPP....K..QLW ....LF..II..TRVI.ERS
Aca.cas.II     .M....TPG.R..VI...M. P.Q.V.AQSNK.V..IV..T LL.GC..TQRD.FIEKDVMM N..MWLTVPT.T..K..QLW .....F..II..NKVL.EE.
Sac.cer.II     .........TR..LSQ.CA. PLQ.VS.QSNK.CM.IV..T LC.IRK.TLRD.FIELDQ.L NM.YWVVIPT...IK...LW S...IL.VAI..NGML.ID.
Sch.pom.II     .M.......TR..IQ.ITM. P.Q.VS.QSNK.VM.IV..T LA.VRKF.LRDNF..RNA.M N.MLWVI.PP.V..K..VLW ....IL..II...GML.E..
Nos.loc.II     ...M...YNSK..LS..... S.H..S.QSNK.VM.II..T LL.VCR.TS.GVFIKR..FC NLVYASANIK...AR..IL. ....LF..T..--KV..MD.
Vai.nec.II     ...M...YNSK..LE..C.. S.QVLS.QSNK.VM.IV..S LTALR.FTLRDSFFDRR.TM QL.YSVI.NF...SY..KLW ....IL.YI....YVI.R..
Pla.fal.II     ...LA..H.TRS.IKH..I. QRQ.VS.QGNK.VM.IV..S LLAIRKFTRRDNF..K...M SL.IWIVIPT...IK.RALW ....IF.ML.Q.GKVI...N
Mas.inv.II     .M...N.L...S.VKN..A. PFQ.V..QKNS.CM.VV..S LL.CS.I.RRD.F..EDVMM NLAL.IKIPQ....R...LW ....AF.ML....VVR..K.
Tri.vag.II     .......QT.RT.V.HI.A. P.Q..S.QANK.V..LV..A LL.CR...KRD.F..RNQ.M NLMMWLI.PP.C..K.VQLW S....F......RHVI.RD.
Try.bru.II     .......LLTK..LI.M.M. P..FVS.NKSA.CM.IV..S LL.S.R.TD.D.F.DKYF.Q SVALWLQ.PI....K.R.LW .....F..I....VVM.RR.
Gia.lam.II     ...AL..L..V..IS..CM. SNQVVSVKDNR..VYIV..V LL.C..FTG.DVTIPFAR.C EYIMWMNVSQ...CY.EARW .....F.M.V.G.VMCGRIA
Sac.cer.III    ......T...R...IN..G. KN.LL..KS.E...AAT..F .T.S..I.H.DSFYDRATLT QL.SMM-IPP...MK.YYLW .....F..LIK.G.VI.RGS
Gia.lam.III    .IFYM.GQ..R...GI..GS .E...S..H.ECM..LT..F LT.I....G.GIFM.RQ.YC QHVSYGC.SQ.C.VY.RQL. S....L.VLMK.DY.I.Q.S
Try.bru.III    .V.FV.T.K.R...LQ..ST AR...SAKN.E...ACT..F LAA...VTSRDVFFDRG.FS QMVSHWR.PI....K.VELW ....LFE.IVREGYVAFLDS
Mus.mus.I      .A.F....LGR...YV.ACT DQQYLV.KD.Q.LA.LI..H MVSGANMTIRGCFF.R.QYM ELVYRG-TFP....K.F.LW .....V.TL.IE.QVI.RE.
Dro.mel.I      .A.Y....V.R...YN.VN. AS.YLV.KD.T.LG.LI..H VISGVK..IRGRFFNR.DYQ QLVFQG-SLP.T..K.AVLW S...IL.TIIIE.QVQ.R..
Try.bru.I      .V.....I.TR..VET..DA NI.YLV.TS.R..R.LI..H VAAGV.VTLRDKFFDHSTFV QLVYNGLIPI....M.R.MW ....L..VMVR..TVMFA.S
Sac.cer.I      .M.F..N.N.R...LN.ANT DSQYL..TS.S.VR.LI..H ..AGVW.TS.DSFF.R.QYQ QYIYGCVTLP.T.FK.Y.LW ....I.TTV.LENEVLF.D.
Sch.pom.I      .M.F...TN.RS..QFIANT DSQYLV.TS.D.LR.LI..H VVMGVW.TC.D.FYIRD.YQ QL.FQAKTLP...QR.GI.W ....I..SV.LEGSVLFDD.
```

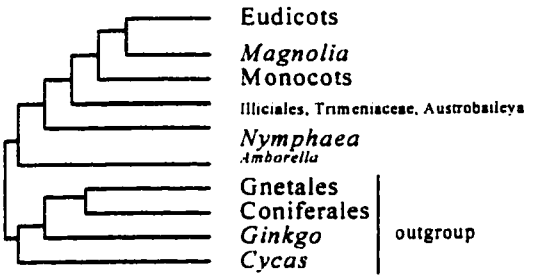
```
Sulf.acido    LLLEGVFDKKAIGNQPESNL HWSIREYGTEYGKWLMDNVF KMFIRFLEHRGFTMTLEDIT IPDEAQNEITTKIKEGYSQV DEYIRKFNEGQLEPIPGRTI
Por.yez.II    E.IT.IV..RTV.SAANGLI .VTWK.K.P.RTCV.ISAIQ VLVNHYVIM..QSIGIG.TI ADAHTDANVRAT.TGAQAE. TQLE.RAQ..E.TLL..KSM
Bon.ham.II    Q.IC.IV..RTV.SSANGLI .ITWK.F.PKITDT.ISQIQ VLVNHYILQ.RQSIGIG.TI AD.ATMRNVIDT.QGAKEE. KLLV..AQ.KE.VLL..KGM
Dro.mel.II    E.IM.ILC..SL.TSAG.L. .ICFL.L.HDIAGRFYG.IQ TVINNW.LFE.HSIGIG.TI ADPQTY...QQA..KAKDD. INV.Q.AHNME...T..N.L
Art.sal.II    E.IM.ILC..TL.AXAG.I. .IIFL.L.HDICGKFXG.IQ TVVNNW.LYE.HSIGIG.TI ADPQTY.S.Q.T..KAKED. IDV.T.AHNNE...T..N.L
Cra.gig.II    M.IS.ILC..TL.TSSG.LA .VVFM...WVIAGEMYGHIQ TLVNNW.LLE.HSIGIG.TI ADPQTYID.QDT..KAKHD. I.V.E.AHNDD...T..N.L
Ily.obs.II    E.IS.ILC..TL.TSAG.LV .IVFL.M.F.VAGE.YG.IQ TVVNNW.LIE.HSIGIG.TI ADQQTYH..QET.RKAKAD. I.V.E.AHNDE...T..N.L
Mus.mus.II    E.IM.ILC..SL.TSAG.LV .I.YL.M.HDITRLFYS.IQ TVINNW.LIE.H.IGIG.SI ADSKTYQD.QNT..KAKQD. I.V.E.AHNNE...T..N.L
Cae.ele.II    E..S.IVCS.TV.KSAGNL. .VVTL.L.Y.IAANFYSHIQ TVINAW.IRE.H.IGIG.TI ADQATYLD.QNT.RKAKQD. VDV.E.AHNDD...T..N.L
Hel.sta.II    K..S.ILC..SL.ASSG.LQ .IIHH.L.SDATADFYAYIQ MVTNHW.LVT.H.IGIA.TI ADAKTYSD.Q.A..KAK.D. V.V.E.AHNDE...M..N.L
Spirogyra.II  E..A.ILC..SL.TSGG.LV .IIWE.V.PDAARKFLGHTQ WLVNYW.LQQ..SIGIG.TI ADASTMDT.NET.AKAKTE. KDL.EAAC.K...AQ....L
Arau.het.II   E..T.TLC..TL.TSGG.LI .VIWE.V.PDAARKFLGHTQ WLVNYW.LQH..SIGIG.TI ADSATMEK.NET.AKAKQD. K.L.KASQ.KS...Q....L
Cyc.rev.II    EV.S.TLC..TL.TSSG.LI .VIWE.V.PDAARKFLGHTQ WLVNYW.LQQ..SIGIG.TI ADAATMET.NET.SKAKAE. NQL.QLAHQKA..AE....M
Eph.vir.II    EVIT.TLC..TL.ASSG.LI .VIWE.V.PDAARKFLGHTQ WLVNYW.LQQ..S.GIG.TI ADATTMDT.NET.QDAKIK. Q.L.E.YMAHK..QE....L
Gnk.bil.II    EV.S.TLC..TL.TSSG.LI .VIWE.V.PDAARKFLGHTQ WLVNYW.LQQ..SIGIG.TI ADAATMET.NET.SKAKNE. KQL.KAAQ.KA..AE....M
Mag.sou.II    E..A.TLC..TL.TSTG.LI .VIWE.V.PDAARKFLGHTQ WLVNYW.LQN..SIGIG.TI ADASTMEK.NET.SKAKNE. K.L.KAAQ.K...AE....M
Nym.odo.II    E..S.TLC..TM.TSSG.LI .VIWE.V.PDAARKFLGHTQ WLVNYW.LQNA.SIGIG.TI ADASTMEK.NET.SKAKNE. K.L.KAAQ.K...AE....M
Zea.mays.2    E..S.TLC..SL.TGSG.LI .VIWE.V.PDAARKFLGHTQ WLVNYW.LQN..SIGIG.TI ADASTMET.NDT.SKAKNA. K.L.K.AH.K...AE....M
Arab.tha.2    E..A.TLC..TL.TSNG.LV .VIWE.V.PDAARKFLGHTQ WLVNYW.LQN...IGIG.TI ADSSTMEK.NET.SNAKTA. KDL..Q.QGKE.D.E....M
Dic.dis.II    E..A.ILC.RSL.AANG.II .VVMN.H.HDTCRLFI.QTQ TVVNHW.IN.....GIG.TI ADSATMAKV.LT.SSAKN.. K.L.I.AQNK.F.CQ..KSV
Aca.cas.II    E.VS.ILN..TL.TSHK.LV .VIWN.H.S.VC.HFLNQ.Q HVVNYW.L.H..SVGVG.TI ADE.TLAK..QT.RKAKDE. K.RQLEAQQ....RQ....M
Sac.cer.II    QIIF..VE..TV.SSNGGLI .VVT..K.PQVCAK.FG.IQ .VVNFW.L.N..STGIG.TI ADGPTMR...ET.A.AKKK. LDVTKEAQANL.TAKH.M.L
Sch.pom.II    EIIY..V...TV.ASQGGLV .TIWK.K.P.IC.GFFNGIQ RVVNYW.L.N..SIGIG.TI ADADTMK.V.RTV..ARR.. A.C.QDAQHNR.K.E..M.L
Nos.loc.II    V..N..IC..SV.TAQG.LI .IIAND..H.EITRFI.SLQ .LISTY.TLIT.SVGIG.TI SSP.TMAH.SRA.GDAKAE. KQL.YDSRRRS..KL..MNL
Vai.nec.II    EI.S.II....V.STQGGLI .IIANDF.PDRVTCFF.DAQ ..MNLYFATIA.SIGIG.AI ADK.TMSQVQRS.ETAKE.. N.I.V.AQKNK..RL..MSM
Pla.fal.II    E..S.IIC.RTV.SSSG.LI .VLWH.M.PDKT.DFLSALQ .VTNNW..YV...VSCS..I ASNKVLGKVREILDKSK.E. SKLVE.AQK.E..CQ..KSL
Mas.inv.II    Q.HT..ITN..V.KSQG.II .ILWKDQ.PMAARDFLSR.Q LLTNAYILT...SVGT..TL ADPDTLQAVKAE.EGAKRN. KIH.DDARA.R.KVQA..SL
Tri.vag.II    N..A.ILG..TVARSEG.LI .VV.NS.N.NIA.DFLNQTQ LIVNNW......SIG.I.CV V..FVLQ.VKHE.DDVE.K. QAT.I.AQD....RM..MSY
Try.bru.II    Q..C.PIT.SIV.AA.G.LI .VIFN.H.SDEVARFING.Q RVTTF..LNF..SVGVQ.TV ADSDTLRQMNDVLVKTRRN. EKIGAAA.NRT.NRKA.M.L
Gia.lam.II    SVVG.-----TESA----LI .ILF.D..I.PCRAFI..CQ RVVC..MLDH..SVGMG.MV SSEHTERKVAEIQTKLSKDI S.LFKLSIHYKIKLA..QSR
Sac.cer.III   QI.S..M..SVL.DGKH.VF YTIL.D..PQEAANA.NRMA .LCA...GN...SIGIN.V. PA.DLKQKKEELVEIA.HKC ..L.TL..K.E..TQ..CNE
Gia.lam.III   EH.V.RLT.TFLASS-NCIF YFLVQN..PVSAARI.LRFA .VAA...MNY...IGID.VM PSQRVLGKKEVIVQ...EKA Q.K..DYES.K..A...S.V
Try.bru.III   CFIS.RL...LL.GGKDGLF ARLHTIA.GG.TARV.SRIA QFTS.Y.TNY..SLG.G.VA PTP.LNKQKAAVLARSVEVC .GL.KSAKT.RMI.L..L.V
Mus.mus.I     E..C..L..AHY.SSAYGLV .CCYEI..G.TSGRVLTCLA RL.TAY.QLY...LGV...L VKPN.DVVRQRI.E.AVKAA LSLPETASCDSKDQRDFNM.
Dro.mel.I     E..V..L..QQY.ATTYGLI .CMYEL..GDVSTL.VTAFT .V.TF..QLE...LGVK..L VT.V.DRKRRKI.R.AVAAA L.LEDEPPHD-KDSKFRVLL
Try.bru.I     E.IT.FMC..QL.ASNM.AP .HVYEL..PHRTGQ.FAAFG RVLLLA.RKE.LSLAMD.MF LV..--ERRCDLLRK----L .DIALDVPDE--------E-
Sac.cer.I     A..C.IL..SQY.ASKYGIV .SLHEV..P.VAAKVLSVLG RL.TNYITA-A..CGMD.LR LTA.GNKWR.DIL.TAAAE. TNLDKDTPAD--DNNKSGIL
Sch.pom.I     E..C.IL..SSF.ASAFGLV .SVHEL..PDIAGR.LSVLS RL.TAYAQM-...CRMD.LR LDEQGD.WRRQLLENAA.EY VGLSTDSPIA--DDEKLQGL
```

```
Sulf.acido    EESLESYILDTLDKLRKVAG EIATKYLDPFNNVYIMAITG ARGSELNITQMT
Por.yez.II    M..F.VEVNKV.NGA.DTS. SS.QLS.LKS..IKR.VSA. SK..FI..S.IC
Bon.ham.II    M..F.TEVNKV.NGA.DKS. AS.QRS.LKS..IKR.VSA. SK..FI..S.IC
Dro.mel.II    RQTF.NKVNRI.NDAHDKT. GS.K.S.TEY..LKA.VVS. SK..NI..S.VI
Art.sal.II    RQTF.NQVNRI.NDA.DKT. GS.KNS.TEY..LKA.VVS. SK..NI..S.VI
Cra.gig.II    RQTF.NMVNRI.NDA.DKT. SK.Q.S.SDY..FKA.VVA. SK..KI..S.VI
Ily.obs.II    RQTF.NQVNRI.NDA.DKT. SK.Q.S.SE...FKA.VVA. SK..KI..S.VI
Mus.mus.II    RQTF.NQVNRI.NDA.DKT. SS.Q.S.SEY..FKS.VVS. .K..KI..S.VI
Cae.ele.II    RQTF.NKVNQI.NDA.DRT. SS.Q.S.SE...FKS.VVS. SK..KI..S.VI
Hel.sta.II    RQTF.NQVNRI.NDA.DKT. SL.Q.S.SE...FKS.VVA. SK.NKI..S.VI
Spirogyra.II  M..F.NRVNQV.N.A.DD.. RA.QSS.SES...KA.VTA. SK..FI..S..I
Arau.het.II   M..F.NKVNQV.NRA.DE.. SS.Q.S.SES...KA.VTA. SK..FI..S..I
Cyc.rev.II    M..F.NRVNQV.N.A.DD.. SS.Q.S.SES..LKA.VTA. SK..FI..S...
Eph.vir.II    L..F.NQVNQV.N.A.DD.. NS.QRS.SES..LKA.VTA. SK..FI..S...
Gnk.bil.II    M..F.NRVNQV.N.A.DD.. SS.QRS.SES..LKA.VTA. SK..FI..S...
Mag.sou.II    M..F.NRVNQV.N.A.DD.. SS.Q.S.SES..LKA.VTA. SK..FI..S...
Nym.odo.II    M..F.NRVNQV.N.A.DD.. SS.QRS.SES..LKA.VTA. SK..FI..S...
Zea.mays.2    M..F.NRVNQV.N.A.DD.. SS.QNS.SES..LKA.VTA. SK..FI..S...
Arab.tha.2    RDTF.NRVNQV.N.A.DD.. SS.Q.S.AET..LKA.VTA. SK..FI..S...
Dic.dis.II    I.TF.QKVNQV.N.A.DT.. SS.QDS.SED..LKA.VTA. SK..FI..S..M
Aca.cas.II    M..F.FV.NQI.N.A.DD.. NS.Q.S.RRS..FKA.V.A. SK..AI..S.VL
Sac.cer.II    R..F.DNVVRF.NEA.DK.. RL.EVN.KDL...KQ.VMA. SK..FI..A..S
Sch.pom.II    R..F.AKVSRI.NQA.DN.. RS.EHS.KDS...KQ.VAA. SK..FI..S..S
Nos.loc.II    Q.T...K.NLA.N.A.DIS. TR.VES.NHL.GLKQ.LKA. SK..YI..S.I.
Vai.nec.II    R..F..QVNYI.N.P.DIS. AS.S.S.SFC..MRT.VLA. SK..FI..S.V.
Pla.fal.II    Y..F.TRVNNE.NCA.EM.. KV.SES..ER..IFS.VAS. SK..II..S.II
Mas.inv.II    V..F.AKTNKS.QDALSN.. KKSLSS.KYD..FKL.IES. SK...M..C.I.
Tri.vag.II    MQ.F.TEVNNLTNEILSKTY KVINAKIRGD.SLSE.LSA. SK.ADT.MS.II
Try.bru.II    LQ.F.ADVNSA.N.C.EE.A KK.LSNVRRT.SFKV.IEA. SK.TD...C.IA
Gia.lam.II    NDAF.QEVISKVSGTSLALE KVI.DAAPHR.ALLV.INA. SK.KKF.MM.IS
Sac.cer.III   .QT..AK.GGL.S.V.EEV. DVCINE..NW.APL...TC. SK..T..VS..V
Gia.lam.III   Q.T..ATLNQI.SNV.ESCA Q..L.E.HFT.KPL..SLC. SK..PI..A..I
Try.bru.III   KQ...ARLNTE.S.V.DEC. TA.VQT.SIH..PL..VQS. SK..A...A..M
Mus.mus.I     DMKFKEEVNHYSNEIN.ACM PLGLHRQF.E..LQM.VQS. .K..TV.TM.IS
Dro.mel.I     DRKYK.LLDGYTNDINSTCL PRGLITKF.S..LQL.VLS. .K..MV.TM.IS
Try.bru.I     -ATAAPM.A.YAT.IQQEFV PQRMLVPF.K.HLLL.T.S. .K..N..A..IS
Sac.cer.I     DAVTS.KVNAITSQVVSKCV PDG.MKKF.C.SMQA..LS. .K..NV.VS.IM
Sch.pom.I     DAAMKGKMNGLTSSIINKCI PDGLLTKF.Y.HMQT.TVS. .K..NV.VS.IS
```

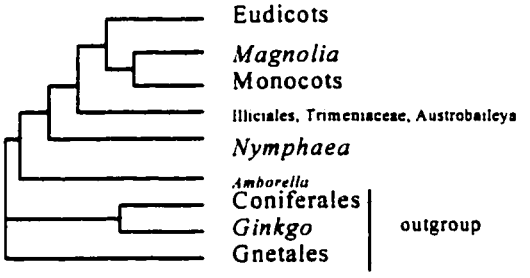
Appendix 5: Results of this study of regions D-F of RPB1 in plants compared with the most recently published studies, which were based upon larger datasets and found that *Nymphaea*, a herbaceous dicot, was amongst the earliest extant angiosperms.



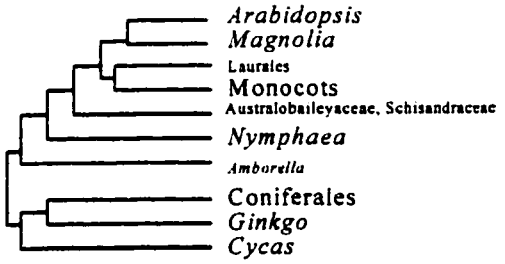
A maximum-likelihood phylogenetic tree of the 278-position amino acid sequence alignment of plant RPB1 sequences, inferred using the JTT model of substitution. The γ parameter α = 1.01, and the fraction of invariable sites = 0.41.



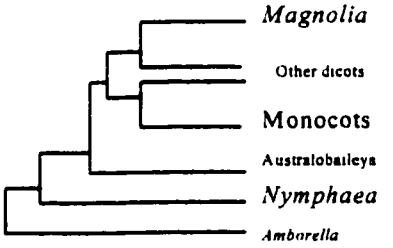
Adapted from: Qiu *et al* (1999) *Nature* 402:404-407. Parsimony (PAUP*) analysis inferred from a 8733 bp alignment of nuclear 18S rRNA, mt *atp1* & *matR*. cp *atpB* & *rbcL* from 105 taxa.



Adapted from: Soltis *et al* (1999) *Nature* 402:402-404. Parsimony (PAUP*) analysis inferred froma 4733 bp alignment of nuclear 18S rRNA and cp *atpB* & *rbcL* from 567 taxa.



Adapted from Parkinson *et al* (1999) *Current Biology* 9:1485-8. ML analysis of 6564 bp alignment of mt (SSU rDNA, *coxI*, rps2), cp rbcL and nuclear SSU rDNA from 51 taxa inferred with fastDNAml, PUZZLE & PAUP* with 4 γ-distributed rates, accounting for rate heterogeneity



Adapted from Mathews & Donoghue (1999) *Science* 286:947-9. Parsimony (PAUP*) analysis of 2208 bp alignment of phytochrome A & C from 26 angiosperms. *PHYA* & *PHYC* used to root one another.