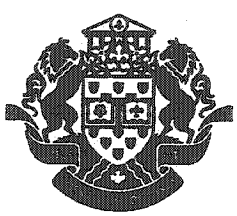


Université d'Ottawa • University of Ottawa

Université d'Ottawa · University of Ottawa

FACULTÉ DE ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Mehrdad HAJIBABAEI

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

Ph.D. (Biology)

GRADE - DEGREE

Department of Biology

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Molecular Evolution of the RNA Polymerase Genes and the Phylogeny of Seed Plants

G. Drouin

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

CO-DIRECTEUR DE LA THÈSE - THESIS CO-SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

G. Carmody

L. Gillespie

S. Graham

D. Hickey

X. Xia

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Molecular evolution of the RNA polymerase genes and the phylogeny of seed plants

Mehrdad Hajibabaei

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
in partial fulfillment of the requirements for the Ph.D. degree in the

Ottawa-Carleton Institute of Biology

Thèse soumise à
Faculté des études supérieures et postdoctorales
Université d' Ottawa
en vue de l'obtention du doctorat

L' Institut de biologie d' Ottawa-Carleton

Your file Votre référence
ISBN: 0-612-89996-9
Our file Notre référence
ISBN: 0-612-89996-9

Table of Contents

List of abbreviations

atpA	ATP synthase alpha subunit
atpB	ATP synthase beta subunit
cox1	Cytochrome c oxidase subunit I
HKY	Hasegawa-Kishino-Yano model of nucleotide substitution
ITS	Internal Transcribed Spacer
Ka	number of nonsynonymous substitutions
Kaa	number of amino acid substitutions
Ks	number of synonymous substitutions
ML	Maximum Likelihood
MP	Maximum Parsimony
NS	Normalized Splits
psaA	photosystem I P700 chlorophyll A apoprotein A1
psbB	thylakoid structural protein
PVP	Poly Vinyl Pyrrolidone
rbcL	ribulose -1,5-bisphosphate carboxylase/oxygenase large subunit
RNAP	RNA polymerase
rpa1	largest subunit of RNA polymerase I
rpb1	largest subunit of RNA polymerase II
rpc1	largest subunit of RNA polymerase III
rRNA	ribosomal RNA
tRNA	transfer RNA

Acknowledgments

I would like to thank Dr. Guy Drouin, my Ph.D. advisor, whose knowledge, expertise and views helped me a lot throughout this project. I really enjoyed working in Dr. Drouin's lab, and I appreciate his approach in supporting and supervising graduate students. I am grateful to other members of my graduate committee: Dr. George Carmody (Biology Department, Carleton University) for his purposefulness at all levels of this work; and Dr. Donal Hickey (Biology Department, University of Ottawa) for his thought-provoking questions and several helpful discussions and guidance during my career as a Ph.D. student. I would also like to acknowledge Dr. Hickey's efforts in providing research facilities, especially computer facilities, for graduate students of the Biology Department.

I would like to thank the past and present members of Drouin's lab including James Nickerson, Jennifer Nickerson, Banoo Malik, Junnan Xia, Robert Morris, Geoffrey Morris and Claude Desruisseaux for collaboration and helpful discussions.

I would like to thank Junnan Xia and Jennifer Nickerson (both M.SC. students, Biology Department, University of Ottawa) for providing the sequence information of *rpb1* genes. I would like to thank Brent Hall (U.C. Santa Cruz, USA), Hugette Allard (Biology Department, University of Ottawa) and Ed Bruggink (Biology Department, Carleton university) for providing plant material.

A special thank you goes out to my friends Robert Beiko and Greg Singer, with whom I had several useful discussions (mostly over coffee and donut breaks) and shared the experience of organizing Ottawa Bioinformatics Discussion Group. I would also like to thank Greg Singer for his help in computer related issues.

I would like to acknowledge NSERC and OGSST for the financial support they provided for my graduate studies.

Finally, I am grateful to my family for their encouragement, support and patience throughout this work and in my entire life.

Abstract

This work is a union of molecular biology and evolutionary biology in addressing questions regarding the evolution of organisms and their genes. This project focuses on major groups of seed plants (angiosperms, and four gymnosperms groups: conifers, cycads, Ginkgo and Gnetales). These organisms diverged roughly three hundred million years ago and evolved independently since then. Therefore, morphological characters have limited power in resolving their phylogeny. With the advancement of molecular methods, sequences of genes have been widely used in addressing evolutionary relationships. Plant molecular phylogenies have been traditionally done using chloroplast gene *rbcL* and nuclear rRNA genes. While these molecular markers revolutionized plant systematics, they do not contain enough phylogenetic information to resolve the phylogeny of seed plants.

I used normalized splits, as a measure from spectral analysis, in examining data sets and identifying genes and sites within genes that are more reliable for building phylogenies and combining data sets. I used the genes encoding the largest subunit of RNA polymerase I, II and III (*rpa1, rpb1* and *rpc1*, respectively) as a new line of evidence from the nuclear genome to address seed plant phylogeny. These protein-coding genes refute the Anthophyte hypothesis (which relates Gnetales and angiosperms) and support the Gnepines hypothesis (which places Gnetales within conifers as a sister group to Pinaceae).

I also examined other available data sets from chloroplast, mitochondrial and 18S rRNA genes. A combined data set of RNA polymerases and seven other genes strongly

support the Gnepines hypothesis. Other branches of seed plant phylogeny are also well supprted using this data set. Angiosperms and gymnosperms form monophyletic groups, cycads are at the base of gymnosperms, followed by Ginkgo and a branch leading to conifers and Gnetales.

I also studied the rate and pattern of molecular evolution of *rpa1, rpb1* and *rpc1* genes. A gene phylogeny of these genes strongly support that they share a common ancestor and that *rpb1* and *rpc1* are closer to one another than either are to *rpa1*. The pattern of functional constraint in these genes clearly follows known structural properties of the RNA polymerases. Evolutionary rate analyses show that *rpa1* evolves more rapidly at non-synonymous sites and at the amino acid level, followed by *rpc1* and *rpb1*. Interestingly, the evolutionary rate of these genes (and the proteins they encode) is positively related to the amount of concerted evolution in the genes transcribed by each RNA polymerase. This suggests a possible impact of concerted evolution (generated by recombination) on the molecular evolution of the largest subunit of RNA polymerases.

Résumé

Cette thèse cherche à répondre à des questions évolutives à l'aide d'outils développés dans les domaines de la biologie moléculaire et de l'évolution moléculaire. En particulier, elle vise l'étude de l'évolution des cinq groupes majeurs de plantes à graines: les angiospermes et les quatre groupes de gymnospermes (les conifères, les cycadales, Ginkgo et les gnétacées). Ces groupes ont eu leur dernier ancêtre commun il y a environ 300 millions d'années, date à partir de laquelle ils ont évolués de façon indépendante. Par conséquent, les caractères morphologiques ne sont que de peu d'utilité pour étudier leurs relations évolutives. Par contre, les progrès récents au niveau des outils de biologie moléculaire et d'évolution moléculaire nous permettent maintenant d'étudier les relations évolutives des espèces à l'aide de séquences de gènes. Les phylogénies moléculaires des plantes sont le plus souvent reconstruites à l'aide des gènes chloroplastiques codant pour le *rbcL* et les gènes nucléaires codant pour l'ARN ribosomal 18S. Quoique ces gènes aient révolutionné la systématique des plantes, ils ne contiennent pas assez d'information phylogénétique pour pouvoir établir les relations évolutives des cinq groupes majeurs de plantes à graines.

J'ai utilisé des divisions normalisées, une mesure d'analyse spectrale, pour examiner et identifier les gènes ainsi que les sites des gènes qui sont le plus appropriés à la reconstruction d'arbres phylogénétiques. J'ai utilisé les gènes codant pour la plus grosse sous-unité des gènes codant pour les ARN polymérases I, II et III (nommés les gènes *rpa1*, *rpb1* et *rpc1*, respectivement) comme de nouvelles séquences de gènes nucléaires codant des protéines pour étudier les relations évolutives des cinq groupes majeurs de

plantes à graines. Les analyses des séquences de ces gènes ne supportent pas l'hypothèse des Anthophytes (qui suppose une relation évolutive récente entre les gnétacées et les angiospermes) mais supporte l'hypothèse des Gnépines (qui suppose une relation évolutive récente entre les gnétacées et les pins).

J'ai aussi analysé les séquences de plusieurs gènes chloroplastiques, mitochondriaux ainsi que celles codant pour l'ARN ribosomal 18S. Les séquences combinées des gènes codant pour la plus grosse sous-unité des ARN polymérases I, II et III ainsi que sept autres gènes supporte l'hypothèse des Gnépines. Les autres branches de cette phylogénie sont aussi fortement supportées. Les angiospermes et les gymnospermes y sont des groupes monophylétique, les cycadales sont retrouvées à la base des gymnospermes, suivies par Ginkgo et une branche menant aux conifères et aux gnétacées

J'ai aussi étudié les taux et types de changements des gènes codant pour la plus grosse sous-unité des ARN polymérases I, II et III. La phylogénie de ces gènes démontre clairement que les gènes *rpb1* et *rpc1* ont une origine commune plus récente que celle du gène *rpa1*. Les contraintes fonctionnelles de ces gènes sont clairement corrélées avec les propriétés structurelles des protéines codées par ces gènes. L'analyse des taux d'évolution de ces gènes démontre que les substitutions nonsynonymes s'accumulent plus rapidement dans *rpa1*, à un taux plus bas dans *rpc1* et a un taux encore plus bas dans *rpb1*. L'ordre des taux d'évolution de ces gènes est positivement corrélé avec la quantité d'évolution concertée observé dans ces gènes. Cette observation suggère que les mécanismes responsables de l'évolution concertée influencent aussi le taux d'évolution des gènes codant pour la plus grosse sous-unité des ARN polymérases I, II et III.

Chapter 1

General Introduction

Diversity and importance of seed plants

Although many seed plant groups are known from the fossil record, only five groups of seed plants survived to the present: angiosperms (with seeds in an ovary) and four gymnosperm groups (with seeds that are not protected by an ovary): conifers, cycads, ginkgo, and Gnetales.

Angiosperms (flowering plants) are the dominant group of plants in most terrestrial habitats. With about 250,000 species they include most of the agriculturally important plants. Conifers (e.g., pine, spruce, larch) are in general needle-leaved, evergreen woody trees and shrubs. With 550 species, they make up the dominant trees in cold forests of Northern Hemisphere and higher latitudes. They are the most economically important group of gymnosperms because they are a major source of wood and resin. Cycads are tropical plants with a single thick stem and large leathery compound leaves. With 150 species they are mainly well known as ornamentals (e.g., *Cycas*). Ginkgo with a single species, *Ginkgo biloba* from China, is a tree with fan

shaped leaves. It is well known as an herbal supplement. Gnetales with 70 species are the least familiar group of seed plants. They include three genera: *Ephedra*, a green-twigged shrub of Western Hemisphere and Eurasian semi-deserts, famous as the source of the medicine pseudoephedrine; *Gnetum* a tropical tree or vine; and *Welwitschia*, with only one species found only in the Namib desert of Southern Africa. This species has only two leaves that grow from the base and get shredded lengthwise by the winds.

Early classifications

Evolutionary relationships of seed plants have long been a matter of debate and controversial theories. Perhaps the most famous saying about this issue is by Charles Darwin who called the emergence and relationships of flowering plants an "abominable mystery" and a "perplexing phenomenon" (Darwin 1903).

Seed plants radiated rapidly very long ago. They appeared in the late Devonian, about 370 million years ago, and at least three of the five groups diverged 290-320 million years ago, within the late Carboniferous (Kenrick and Crane 1997).

The relationships of Gnetales and the origin of angiosperms are the two major issues in the seed plant phylogeny. These questions have been closely tied together in early theories. The Englerian theory (Wettstein 1907) proposed that Gnetales are the sister group of angiosperms. This theory was based on morphological similarities between Gentales and Amentiferae (a group of angiosperms that were considered most primitive by the Englerian theory). The simple flowers of the wind-polinated Amentiferae (unisexual, lacking sepals and petals, having either stamens or carpels) were considered

homologous with the compound flower-like strobili of Gnetales (Doyle 1998a). This view lost its credibility when Magnoliidae (with complex, bisexual flowers) were proposed, based on wood anatomy and pollen structure, as the earliest flowering plants. Another theory was that of Arber and Parkin (1907; 1908), which also associated Gnetales with angiosperms. This theory, however, related the angiosperms and Gnetales with the extinct Mesozoic Bennettitales (with large, often bisexual flowers) implying that the flowers of Gnetales are reduced rather than primitive. Later ideas were centered around the differences of angiosperms and Gnetales and were against grouping them together. For example, Bailey used the differences in origin of vessels to separate angiosperms and Gnetales (Bailey 1944) and xylem features to further relate them with conifers and *Ginkgo* (Bailey 1953). Another view suggested that Gnetales are diphyletic (Eames 1952) and separated *Ephedra* from *Gnetum* and *Welwitschia*. Several authors later investigated homologies of angiosperm organs with fossil groups such as *Caytonia* and *Glossopterids* (Gaussen 1946; Stebbins 1974; Doyle 1978).

Cladistic analyses of morphological data (the Anthophyte hypothesis)

By the mid 1980s a new picture of seed plant phylogeny emerged. Crane (1985) and Doyle and Donoghue (1986; 1992) analyzed all available morphological characters using the principle of parsimony. Theses studies grouped Gnetales and angiosperms with two fossil groups Bennettitales and Pentoxylales. This new grouping was called anthophytes because their common ancestor would have had flower-like structures. Other morphological studies also confirmed the anthophyte grouping (Loconte and Stevenson

1990; Nixon et al. 1994). However, detailed groupings were different between different morphological analyses. For example Crane (1985) directly related Gnetales and angiosperms but Doyle and Donoghue (1992) placed angiosperms basal in the anthophytes.

The anthophyte grouping is based on the similarity in overall organization of reproductive structures. To be exact, there are a number of similar characteristics between these groups: double fertilization (both sperms produced by the male gametophyte fuse with nuclei in the female gametophyte); tunica formation in apical meristem; lignin chemistry; pollen wall structure; lack of archegonia; and conductive vessel-like xylem elements (Doyle 1998b). Although it is controversial which of these characters are real synapomorphies, double fertilization is the most commonly accepted one (Doyle 1998b).

Introduction of molecular phylogenetics (early studies)

Molecular phylogenetic studies of plants, which started more than a decade ago, have revolutionized the field of plant systematics. Several genes, mostly from the chloroplast genome, have been sequenced and used in molecular phylogenies of plants. However, the *rbcL* (large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase) chloroplast gene and the 18S rRNA nuclear ribosomal gene are the most extensively used molecular markers in plant phylogeny. These two genes were chosen mainly because they were easy to sequence.

Using *rbcL*, Chase et al. (1993) analyzed 499 taxa in an extensive study (with the collaboration of 38 authors). This analysis, in which cycads were used as outgroups,

suggested that both Gnetales and angiosperms are monophyletic sister groups and the aquatic plant *Ceratophyllum* is at the base of angiosperms. Another analysis of *rbcL* (Hasebe et al. 1992) with fewer taxa and using translated amino acid sequences and ferns as outgroups confirmed that both Gnetales and angiosperms are monophyletic but contrary to Chase et al. (Chase et al. 1993) this *rbcL* analysis placed Gnetales with gymnosperms.

Further analyses based on 18S rRNA genes concluded that both angiosperms and Gnetales were monophyletic but the relationship between them was not strongly established (Hamby and Zimmer 1992; Doyle, Donoghue, and Zimmer 1994). Hamby and Zimmer's (1992) study, which used 60 taxa rooted with fern species (*Psilotum* and *Equisetum*), did not group angiosperms and Gnetales in maximum parsimony analyses but grouped them in neighbor-joining analyses. The study of Doyle, Donoghue, and Zimmer (1994) also reached different groupings based on the number of taxa and the rooting used. Another study using 18S rRNA grouped Gnetales with conifers (Chaw et al. 1997).

Studies using other, less extensive, sequence data sets including 28S rRNA (Stefanovic et al. 1998), chloroplast ITS (Goremykin et al. 1996) and mitochondrial *cox1* (Bowe and dePamphilis 1997) did not group angiosperms and Gnetales. In fact, *cox1* and ITS sequences supported the grouping of Gnetales and conifers.

The above mentioned molecular studies raised serious doubts about the Anthophyte hypothesis but their results were inconclusive because they had low statistical support and lacked congruence among different genes and methods of analysis (Donoghue and Doyle 2000). Most of these analyses were based on the parsimony

criterion, which is known to falsely group the long branches of a tree (Felsenstein 1978). In fact, some of the above mentioned studies reported different groupings when different taxa were used as outgroups (Doyle, Donoghue, and Zimmer 1994).

New line of molecular phylogenies and the rejection of Anthophyte hypothesis

A number of molecular phylogenetic studies, starting in late 1999, have offered a new picture of seed plant phylogeny (all within the lifetime of this PhD project). By using larger combined sequence data sets, and by applying more rigorous and advanced phylogenetic methods, these new studies addressed the weaknesses of early molecular phylogenies. They refuted the Anthophyte hypothesis, proposed new groupings and resolved several branches. In this section I review these studies (in a chronological order) and their results, focusing on points of agreements and conflicts. Hansen et al. (1999) combined a long sequence of chloroplast DNA (containing protein coding genes, tRNA and ribosomal RNA genes) and nuclear 18S rRNA in a six taxon tree. They only used first and second codon positions of protein coding genes in order to eliminate the effect of saturation of third codon positions. They reached a single topology and high bootstrap support using three methods (maximum parsimony, maximum likelihood and neighbor joining). The major results of their study were the refutation of the Anthophyte hypothesis and the grouping of *Gnetum* with *Pinus* instead of angiosperms. Although this was the first molecular study that firmly refuted the Anthophyte hypothesis, a major

drawback of this study is their limited number of taxa (only *Gnetum* as a representative of Gnetales, and only *Pinus* as the other gymnosperm taxa).

Winter et al. (1999) used nuclear MADS box genes. These genes encode transcription factors which have a role in regulating development (Theissen, Kim, and Saedler 1996). The phylogenetic tree for the translated amino acid sequences of five MADS box genes, reconstructed using the neighbor joining method, refuted the Anthophyte hypothesis. All five genes in this tree supported the grouping of *Gnetum* (the only representative of Gnetales in this tree) and conifers. Furthermore, three of these genes gave high bootstrap support (e.g., 100% or 98%) to this grouping. This analysis, however, could not resolve the relationships of seed plants because it did not include the two other seed plant groups (cycads and *Ginkgo*).

The most recent hypothesis in seed plant phylogeny was proposed by Chaw et al. (2000) and supported by Bowe et al. (2000) (both published in the same issue of Proc. Natl. Acad. Sci. USA) exclusively based on molecular markers. This hypothesis, called the Gnepines hypothesis, places the Gnetales as the sister group of Pinaceae within conifers. Chaw et al. (2000) used mitochondrial small subunit rRNA genes (mtSSU rRNA) and nuclear small subunit rRNA genes (nuSSU rRNA) in combination with *rbcL* genes for all groups of seed plants. Their mtSSU rRNA dataset supported the grouping of Gnetales and Pinaceae in maximum likelihood and maximum parsimony methods with high bootstrap support (98% and 97% respectively). nuSSU rRNA, however, did not support the Gnepines grouping. This data set grouped the Gnetales as a sister group to all conifers. The statistical support for this grouping was high as well (bootstrap support of 91% and 90% for maximum likelihood and maximum parsimony, respectively). The *rbcL*

data set showed three different positions for Gnetales based on the choice of character set or method of analysis (i.e., all codon positions, rate variation among sites, and excluding transitions of third base of codons). The Gnepines grouping was only supported in *rbcL* trees when transitions of third codon positions were excluded from the analysis, and this grouping was supported by average bootstrap values (68% and 64% in maximum likelihood and maximum parsimony, respectively). Chaw et al. (2000) also analyzed a combined data set of the above mentioned genes and the result was in favor of the Gnepines grouping using both maximum likelihood and maximum parsimony (bootstrap support of 100% and 87% for maximum likelihood and maximum parsimony, respectively). Bowe, Coat, and dePamphilis (2000) also used two mitochondrial genes, *cox1* and *atpA*, together with known sequences of *rbcL* and 18S rRNA in a data set containing all groups of seed plants. Their maximum likelihood, maximum parsimony and neighbor joining trees supported the Gnepines hypothesis. However, the nodes connecting Gnetales and Pinaceae were not supported by high bootstrap support (i.e., >85%) in all of their trees.

The Gnepines hypothesis raised a lot of criticism in plant systematics community (Donoghue and Doyle 2000) and has been tested and challenged by recent phylogenetic studies (Sanderson et al. 2000; Rydin, Källersjö, and Friis 2002; Schmidt and Schneider-Poetsch 2002; Soltis, Soltis, and Zanis 2002). However, none of these studies, could resolve the position of Gnetales. Rydin, Källersjö, and Friis's (2002) study, which rejected the Gnepines hypothesis, was based on a parsimony analysis of *rbcL*, *atpB*, 18S and 26S rRNA genes. This work is interesting since it argued that third codon positions of chloroplast genes contain most of the phylogenetic information and should be used in

the phylogeny. They also suggested that there is a conflict between transitions and transversions within each codon position. However, Rydin, Källersjö, and Friis (2002) did not reach a definitive conclusion about the position of Gnetales. Their trees rejected the Gnepines hypothesis and placed Gnetales either as the sister group to all other seed plants or to angiosperms (based on the position of the root). In contrast to Rydin, Källersjö, and Friis (2002), the study of Soltis, Soltis, and Zanis (2002) generally supported the Gnepines hypothesis. These authors used eight genes in maximum likelihood, maximum parsimony and Bayesian phylogenetic methods. They obtained different topologies based on the choice of data set or the method of analysis. Interestingly, this study suggested that most of the discrepancies in seed plant phylogenies involve the third codon positions of chloroplast genes. In an analysis of nuclear encoded phytochrome genes, Schmidt and Schneider-Poetsch (2002), rejected the Gnepines hypothesis, but did not reach a conclusion on the position of Gnetales.

The Gnepines hypothesis was also challenged by structural evidence from the chloroplast genome. In land plants, the chloroplast genome has a conserved structure which contains two copies of an inverted repeat about 10-25 kbp long (Palmer and Stein 1986). However, all conifers have only one copy of this inverted repeat whereas two copies are present in the chloroplast genome of Gnetales (Raubeson and Jansen 1992). If the Gnepines hypothesis is correct then either Pinaceae and all other conifers lost one copy of the inverted repeat (but not the Gnetales) or the second inverted repeat was regained in Gnetales after their divergence from Pinaceae.

Chapter four addresses the phylogeny of seed plants and the position of Gnetales. I used RNA polymerase genes and seven other genes and a combination of different molecular phylogenetic methods to reconstruct the phylogeny of seed plant.

Phylogenetic studies on the origin of Angiosperms

Based on various characteristics, several plant groups have been proposed as being the most primitive of the angiosperms. The most important ones include: Magnoliales (Donoghue and Doyle 1989), Calycanthales (Cronquist's subclass Magnoliidae) and water lilies (Loconte 1996). Early molecular studies, based on *rbcL* and 18S rRNA genes, were helpful in resolving the major groups of angiosperms, but they were inconclusive and lacked high statistical support (Kuzzof and Gasser 2000). However, in 1999 five molecular multigene studies reported a new consensus for the branching order of early angiosperms (Matthews and Donoghue 1999; Parkinson, Adams, and Palmer 1999; Qiu et al. 1999; Soltis, Soltis, and Chase 1999; Graham and Olmstead 2000). These studies employed more advanced phylogenetic methods to analyze sequences of genes from chloroplast, mitochondria and nucleus for up to 560 species. They all supported *Amborella*, a shrub from New Caledonia, as being the single sister taxon to all other angiosperms. Following *Amborella,* Nymphaeales (water lilies), and the ITA clade (Illiciaeae, Schisandraceae, Trimeniaceae and Austrobaileyaceae) were supported as sister to the other angiosperms. This grouping of five early branching extant angiosperm taxa was called the ANITA clade (Qiu et al. 1999). However, a study by Barkman et al. (2000) suggested that a branch consists of *Amborella* + Nymphaeales should be considered the sister group to all other angiosperms. More recent molecular

phylogenetic analyses further confirmed the position of *Amborella* as the sister group to all other angiosperms (Qiu et al. 2001; Zanis et al. 2002). I have touched on this issue in chapter three. Using a property of the spectral analysis (see below) I study the effect of misleading data partitions on an exemplar phylogeny of the basal angiosperms.

Organization and structure of RNA polymerases

RNA polymerase (RNAP) is the enzyme responsible for transcribing genes to RNA in all cells. In bacterial and archaeal cells, a single RNAP produces all RNA types, whereas in eukaryotes there are three nuclear RNAPs responsible for transcribing different genes: RNAP I transcribes rRNA genes; RNAP II transcribes pre-messenger RNA (protein-coding genes); and RNAP III transcribes tRNA, 5S RNA and small RNA genes (Archambault and Friesen 1993). RNAPs are complex multi-subunit enzymes (up to 0.6 MDa). Table 1 compares the subunits of RNAPs. Each RNAP consists of a homologous core structure containing five subunits: largest subunit (RPA1, RPB1 and RPC1 in RNAP I, II and III respectively; and β' in prokaryotic RNAP) and second largest subunit (RPA2, RPB2 and RPC2 in RNAP I, II and III respectively and β in prokaryotic RNAP) form the central mass of the enzyme. The third and fourth core subunits (AC40 and AC19 in RNAP I and III, RPB3 and RPB11 in RNAP II, and α and α' in prokaryotes) are involved in RNAP assembly and the fifth core subunit (RPB6 common in RNAP I, II and III and, ω in prokaryotes) stabilizes the large subunits (Cramer 2002). The core contains functional elements of the enzyme, including the active site, which is in the

largest subunit of the enzyme. A series of recent studies have helped in resolving the structure of the bacterial RNAP (in *Thermus aquaticus*) as well as eukaryotic RNAP II (in the yeast species *Sacchromyces cerevisiaea*) with high resolution (Fu et al. 1999; Zhang et al. 1999; Cramer et al. 2000; Campbell et al. 2001; Cramer, Bushnell, and Kornberg 2001; Gnatt et al. 2001). These studies have shown that there is even a higher degree of similarity, beyond the primary sequence, in the structural properties of prokaryotic and eukaryotic RNAPs. This suggests a similar mechanism for the action of all RNAPs (Ebright 2000).

The largest subunit of RNAPs are approximately 1500-1700 amino acids long and share 8 conserved regions (A-H). Figure 1 shows a schematic diagram of the largest subunit with conserved blocks A-H (information retrieved from Cramer, Bushnell, and Kornberg 2001). The three metal binding aspartate (D) amino acids (shown in red color in Figure 1) in the invariant string "NADFDGD" form the active site of the enzyme (Cramer, Bushnell, and Kornberg 2001).

Molecular evolution of the genes encoding the largest subunit of RNAPs

Genes encoding the largest subunit of nuclear RNAPs in eukaryotes evolved through duplications from an ancestral gene after the divergence of prokaryotes and eukaryotes (Iwabe et al. 1991). There are mixed views on the evolutionary relationships of three eukaryotic RNAPs and the prokaryotic ones. Memet et al. (1988) showed that genes encoding the largest subunits of three RNAPs in eukaryotes share a common ancestor and that *rpb1* and *rpc1* are closer. Iwabe et al. (1991) refuted this view and

suggested that although these genes share a common ancestor, *rpa1* is closer to *rpb1*. A more recent study by Klenk et al. (1995) did not reach a consensus on the branching order of these genes, but generally supported the closeness of *rpb1* and *rpc1*. However, these studies were done with a limited number of genes, especially *rpc1* and *rpa1*. As for the rate of evolution, Iwabe et al. (1991) suggested that *rpa1* evolves faster than *rpb1* and *rpc1*. In chapter five I reconstructed the phylogenetic relationships of *rpa1*, *rpb1* and *rpc1*, and studied the evolutionary rate and the pattern of functional constraint of these genes.

Utility of RNA polymerases as phylogenetic markers

RNA polymerases have been used in several phylogenetic studies. Originally they were used, as an alternative to the popular rRNA molecules, to address the early branching pattern of eukaryotes, bacteria and archaea (Zillig et al. 1989; Iwabe et al. 1991; Klenk, Palm, and Zillig 1994; Sidow and Thomas 1994). Other phylogenetic studies using RNAPs include a wide range of organisms, for example: protists (Klenk et al. 1995; Hirt et al. 1999; Dacks et al. 2002); fungi (Liu, Whelen, and Hall 1999); plants (Denton, McConaughy, and Hall 1998; Oxelman and Bremer 2000); arthropods (Shultz and Regier 2000). Recently, our lab has shown that *rpb1* is a useful phylogenetic marker for plants (Nickerson and Drouin in press).

The utility of a gene in organism phylogeny is usually assessed by factors such as accessibility through molecular methods (e.g., PCR cloning), having evolved through

speciation (being orthologues; to reflect an accurate estimate of the phylogeny of species), containing enough phylogenetic informative sites for a certain divergence level to reconstruct a fully resolved and statistically sound phylogeny (for example a highly conserved gene is not able to discriminate among closely related species), being less prone to phenomena which influence the evolutionary history of genes (e.g., gene conversion). However, no objective, clear-cut criteria exist for choosing genes for phylogenetic analysis of a set of taxa. Genes encoding the largest subunit of RNA polymerases possess the above qualities for a variety of divergence levels and were even proposed as potential universal phylogenetic markers (Nickerson and Drouin in press; Klenk, Palm, and Zillig 1994).

In recent years, with the advancement of sequencing technology and phylogenetic methods, phylogenetic problems are mostly being addressed using a multigene approach (e.g., Qiu et al. 1999; Soltis, Soltis, and Chase 1999). A single gene usually does not contain enough phylogenetic signal to fully resolve a tree, and the result from a single gene is not considered very reliable (low statistical support, sampling error). It is also important to obtain evidence for a phylogeny from different genomic compartments. Plant molecular phylogeny (as was mentioned earlier) has been mostly done using chloroplast and rRNA genes. As a new line of evidence from protein-coding nuclear genes, I have used sequences of *rpa1, rpb1* and *rpc1* to address seed plant phylogeny. I also compared and combined these genes with other available genes (from chloroplast, mitochondria and nuclear ribosomal gene 18S rRNA). Although *rpb1* has been used in several phylogenetic studies, the sequences of *rpa1* and *rpc1* are being used as phylogenetic markers for the first time.

Evolution of Molecular Phylogenetics

Molecular phylogenetics is the use of molecular sequence (DNA, amino acids) information to reconstruct the evolutionary relationships of the molecules. This relationship can be a reflection of the evolutionary history of the organisms carrying the molecules (when the genes are orthologous, meaning that they have passed to each lineage as a result of speciation). Therefore, molecular phylogenies have been widely used to clarify the evolutionary history of different organisms. This field has progressed a lot in recent years due to a parallel progress of DNA sequencing technology, computer science, mathematical biology and molecular evolutionary biology. In fact, the number of publications reporting molecular phylogenies has been increasing exponentially since 1981 (Pagel 1999). There are several books and review articles on how to build phylogenetic trees (i.e., Hillis, Moritz, and Mable 1996; Page and Holmes 1998; Graur and Li 2000; Nei and Kumar 2000), here I only briefly introduce major approaches that I have used in this work.

Maximum Parsimony

Since its introduction, by Edwards and Cavalli-Sforza, 40 years ago (Edwards and Cavalli-Sforza 1963), maximum parsimony (MP) has been the most widely used method for building phylogenetic trees. This method is based on the principle of parsimony, a minimalist principle sometimes also referred to as Ockham's razor. It states that simpler explanations (requiring fewer assumptions) are preferred over more complex, *ad hoc* ones. In a molecular phylogenetic context, this method works by seeking the tree that requires the fewest number of changes along the sites of aligned sequences. This tree is

considered the most parsimonious or the simplest (optimal) description of the data. The simplicity of MP made it a popular method in building phylogenetic trees. However, MP works under the condition that sequences have not undergone (or minimally undergone) multiple substitutions. This condition is not true for most real sequences. Felsenstein's classic and much-cited paper (Felsenstein 1978) showed that when there are multiple substitutions in a site, MP fails to choose the right tree. This problem is also widely known as long branch attraction, meaning that species that have undergone a large number of substitutions (long branches) are often grouped together, despite their different evolutionary histories (Felsenstein 1978).

Maximum Likelihood

The maximum likelihood (ML) approach in building phylogenies is the leading alternative to MP. ML is based on the concept of likelihood of a hypothesis. The likelihood of the hypothesis *H,* given data *D* and a specific model, is proportional to *P*(*D* | *H*), the conditional probability of observing *D* given that *H* is correct (Edwards 1972). In a molecular phylogenetic context *D* is usually the set of sequences being compared and *H* is a phylogenetic tree. The model is usually a stochastic process for substitutions. However, to calculate the likelihood of a tree, there are other parameters (regarding the evolution of D) that we need to either specify or place some prior distributions on them (Steel and Penny 2000). Well known examples of these parameters include parameters associated with the substitution matrix (e.g., transition/transversion bias), and parameters that describe how rates vary across sites (e.g., gamma distribution), and base composition. These parameters typically are either selected to maximize the likelihood or estimated

directly from the data. Although ML is often considered the superior method for building phylogenies, there are some concerns regarding the validity of the stochastic model used and the computational complexity of ML (Steel and Penny 2000).

Spectral Analysis

Spectral analysis (Hendy and Penny 1993) is a method for exploring signals (patterns) in the sequence data independent of building a phylogenetic tree. This method is based on the concept of splits (bipartitions), which means ways to group the sequences (taxa) in two groups. For n sequences (taxa) there are $2^{(n-1)}$ splits. For example, there are 8 splits for a set of 4 taxa. One of these splits groups all four sequences together and three of these splits only separate a single taxa from others, therefore these are not useful for phylogenetic inference. The remaining four splits can resolve groupings for a fully bifurcating tree of 4 taxa. Ideally, sites of a gene (alignment) should all support the same splits. However, this is usually not the case and we often find contradictory signals in the spectrum of a gene. One reason for this contradictory result is the multiple substitutions in a site. To minimize this effect, Hendy and Penny (1993) introduced a mathematical transformation of the data, called Hadamard conjugation, that can correct for the unobserved multiple substitutions.

In a spectral analysis framework, a data set with a strong phylogenetic signal produces the minimum number of splits that are necessary for fully resolving a tree for a finite number of taxa (ideally t-2 splits are necessary to resolve a fully bifurcating tree for t taxa). Therefore, the number of splits produced by a dataset divided by the length of the alignment can be used to estimate the amount of random noise in a set of aligned

sequences (Lento et al. 1995). I used this normalized splits (NS) as a measure to evaluate the noise contained in genes and gene partitions. This work is detailed in chapter three and is also used for evaluating 10 genes and their partitions in seed plants phylogeny (chapter four).

Spectral analysis has been used to evaluate the signal for specific phylogenetic hypotheses, represented by a split (Lento et al. 1995; Lockhart, Penny, and Meyer 1995). With this method one can obtain frequency of support for different groupings in a data set. The numerical value for frequency of support is calculated by dividing the sum of occurrences for a split by the total number of sites in the alignment (Lento et al. 1995). Conventional tree building methods, however, usually produce a single tree as the final result. Spectral analysis allows us to explore the signals that might not be shown in a phylogenetic tree. I have used this approach in evaluating the relative position of Gnetales to conifers in chapter four.

Distance methods

Another major category of phylogeny reconstruction methods are distance-based methods (e.g. Neighbor-joining). These methods are based on calculating distances using a model of substitution. Distance based methods are usually faster than other methods (e.g. ML) and are used for resolving trees with large number of taxa, where computation time is an important factor. However, in this work I have used ML and MP instead of distance-based methods. ML is considered more robust than distance-based methods and MP has been often used in plant phylogenetics.

References

Arber, E., and J. Parkin. 1907. On the origin of angiosperms. J. Linn. Soc. Bot. **38**:29-80.

Arber, E., and J. Parkin. 1908. Studies on the evolution of the angiosperms.The relationship of the angiosperms to the Gnetales. Ann. Bot. **22**:489-515.

Archambault, J., and J. D. Friesen. 1993. Genetics of eukaryotic RNA polymerases I, II, and III. Microbiol. Rev. **57**:703-724.

Bailey, I. 1944. The development of vessels in angiosperms and its significance in morphological research. Am. J. Bot. **31**:421-428.

Bailey, I. 1953. Evolution of the tracheary tissue of land plants. Am. J. Bot. **40**:4-8.

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

Bowe, L., and C. dePamphilis. 1997. Conflict and congruence among three plant genetic compartments: phylogenetic analysis of seed plants from gene sequences of *cox1*, rbcL and 18S rDNA. Am. J. Bot. **84(Suppl.)**:178-179.

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

Campbell, E. A., N. Korzheva, A. Mustaev, K. Murakami, S. Nair, A. Goldfarb, and S. A. Darst. 2001. Structural mechanism for rifampicin inhibition of bacterial RNA polymerase. Cell **104**:901-912.

Chase, M., D. Soltis, R. Olmstead, D. Morgan, and D. Les, et al. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. Ann. Mo. Bot. Gard. **80**:526-580.

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

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

Cramer, P. 2002. Multisubunit RNA polymerases. Curr. Opin. Struc. Biol. **12**:89-97.

Cramer, P., D. Bushnell, and R. Kornberg. 2001. Structural basis of transcription: RNA polymerase II at 2.8 Å resolution. Science **292**:1863-1876.

Cramer, P., D. A. Bushnell, J. Fu, A. L. Gnatt, B. Maier-Davis, N. E. Thompson, R. R. Burgess, A. M. Edwards, P. R. David, and R. D. Kornberg. 2000. Architecture of RNA polymerase II and implications for the transcription mechanism. Science **288**:640-649.

Crane, P. 1985. Phylogenetic analysis of seed plants and the origin of angiosperms. Ann. Mo. Bot. Gard. **72**:716-793.

Dacks, J., A. Marinets, W. F. Doolittle, T. Cavalier-Smith, and J. Logdson. 2002. Analyses of RNA polymerase II genes from free-living protists: phylogeny, long branch attraction, and eukaryotic big bang. Mol. Biol. Evol. **19**:830-840.

Darwin, C. 1903. Pp. 239-240, 220-222 *in* F. Darwin, and A. C. Seward, eds. *More Letters from Charels Darwin*. Appleton, New York.

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

Donoghue, M. J., and J. A. Doyle. 2000. Seed plant phylogeny: Demise of the anthophyte hypothesis? Curr Biol **10**:R106-109.

Donoghue, M. J., and J. A. Doyle. 1989. Phylogenetic analysis of angiosperms and the relationships of Hamamelidae *in* P. R. Crane, and S. Blackmore, eds. Evolution, Systematics, and Fossil History of the Hamamelidae. Clarendon, Oxford, UK.

Doyle, J. 1978. Origin of angiosperms. Annu. Rev. Ecol. Syst. **9**:365-392.

Doyle, J., and M. Donoghue. 1992. Fossils and seed plant phylogeny reanalyzed. Brittonia **44**:89-106.

Doyle, J., and M. Donoghue. 1986. Seed plant phylogeny and the origin of angiosperms: an experimental cladistic approach. Bot. Rev. **52**:321-431.

Doyle, J., M. Donoghue, and E. Zimmer. 1994. Integration of morphological and ribosomal RNA data on the origin of angiosperms. Ann. Mo. Bot. Gard. **81**.

Doyle, J. A. 1998a. Phylogeny of vascular plants. Ann. Rev. Ecol. Syst. **29**:567-599.

Doyle, J. A. 1998b. Molecules, morphology, fossils, and the relationship of angiosperms and Gnetales. Mol. Phylogenet. Evol. **9**:448-462.

Eames, A. 1952. Relationships of the Ephedrales. Phytomorphology **2**:79-100.

Ebright, R. 2000. RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. J. Mol. Biol. **304**:687-698.

Edwards, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge, England.

Edwards, A. W. F., and L. L. Cavalli-Sforza. 1963. The reconstruction of evolution. Heredity **18**:533; Ann. Hum. Genet. 527:104–105.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**:401-410.

Fu, J., A. Gnatt, D. Bushnell, G. Jensen, N. Thompson, R. Burgess, P. David, and R. Kornberg. 1999. Yeast RNA polymerase II at 5 Å resolution. Cell **98**:799-810.

Gaussen, H. 1946. Les Gymnospermes, actuelles et fossiles. Trav. Lab. For. Toulouse Tome II Etud. Dendrol. **1**:1-26.

Gnatt, A., P. Cramer, J. Fu, D. Bushnell, and R. Kornberg. 2001. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Å resolution. Science **292**:1876-1882.

Goremykin, V., V. Bobrova, J. Pahnke, A. Troitsky, A. Antonov, and W. Martin. 1996. Noncoding sequences from the slowly evolving chloroplast inverted repeat in addition to rbcL data do not support gnetalean affinities of angiosperms. Mol. Biol. Evol. **13**:383-396.

Graham, S. W., and R. G. Olmstead. 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. Am. J. Bot. **87**:1712-1730.

Graur, D., and W. H. Li. 2000. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, MA.

Hamby, R., and E. Zimmer. 1992. Ribosomal RNA as a phylogenetic tool in plant systematics. Pp. 50-91 *in* P. Soltis, D. Soltis, and Doyle JJ, eds. Molecular Systematics of Plants. Chapman & Hall, New York.

Hansen, A., S. Hansmann, T. Samigullin, A. Antonov, and W. Martin. 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., R. Kofuji, M. Ito, M. Kato, K. Iwatsuki, and K. Ueda. 1992. Phylogeny of gymnosperms inferred from rbcL gene sequences. Bot. Mag. Tokyo **105**:673-679.

Hendy, M. D., and D. Penny. 1993. Spectral analysis of phylogenetic data. J. classification **10**:5-24.

Hillis, D. M., C. Moritz, and B. K. Mable. 1996. Molecular systematics. Sinauer Associates, Sunderland, MA.

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

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

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

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

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

Kuzzof, R. K., and C. S. Gasser. 2000. Recent progress in reconstructing angiosperm phylogeny. Trends Plant Sci. **5**:330-336.

Lento, G. M., R. E. Hickson, G. K. Chambers, and D. Penny. 1995. Use of spectral analysis to test hypotheses on the origin of pinnipeds. Mol. Biol. Evol. **12**:28-52.

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

Lockhart, P. J., D. Penny, and A. Meyer. 1995. Testing the phylogeny of swordtail fishes using split decomposition and spectral analysis. J. Mol. Evol. **41**:666-674.

Loconte, H. 1996. Comparison of alternative hypotheses for the origin of angiosperms. Pp. 267-285 *in* D. W. Taylor, and L. J. Hickey, eds. Flowering Plant Origin, Evolution and Phylogeny. Chapman & Hall, N.Y.

Loconte, H., and D. Stevenson. 1990. Cladistics of the Spermatophyta. Brittonia **42**:197–211.

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

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

Nei, M., and S. Kumar. 2000. Molecular evolution and phylogenetics. Oxford University Press, Oxford.

Nickerson, J. and G. Drouin. in press. The sequence of the largest subunit of RNA polymerase II is a useful marker for inferring seed plant phylogeny. Mol. Phylogenet. Evol.

Nixon, K., W. Crepet, D. Stevenson, and E. Friis. 1994. A reevaluation of seed plant phylogeny. Ann. Mo. Bot. Gard. **81**:484-533.

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

Page, D. M., and E. C. Holmes. 1998. Molecular evolution: a phylogenetic approach. Blackwell Science Inc., Oxford.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature **401**:877-884.

Palmer, J. D., and D. B. Stein. 1986. Conservation of chloroplast structure among vascular plants. Genetics **10**:823-833.

Parkinson, C. L., K. L. Adams, and J. D. Palmer. 1999. Multigene analyses identify the three earliest lineages of extant flowering plants. Curr. Biol. **9**:1485-1488.

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

Qiu, Y. L., J. Lee, B. A. Whitlock, F. Bernasconi-Quadroni, and O. Dombrovska. 2001. Was the ANITA rooting of the angiosperm phylogeny affected by long-branch

attraction? Amborella, Nymphaeales, Illiciales, Trimeniaceae, and Austrobaileya. Mol. Biol. Evol. **18**:1745-1753.

Raubeson, L. A., and R. K. Jansen. 1992. A rare chloroplast-DNA structural mutaion is shared by all conifers. Biochem. System. Ecol. **20**:17-24.

Rydin, C., M. Källersjö, and E. M. Friis. 2002. Seed plant relationships and the systematic position of Gnetales based on nuclear and chloroplast DNA: Conflicting data, rooting problems, and the monophyly of conifers. Int. J. Plant Sci. **163**:197-214.

Sanderson, M. J., M. F. Wojciechowski, J. M. Hu, T. S. Khan, and S. G. Brady. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol. Biol. Evol. **17**:782-797.

Schmidt, M., and H. A. W. Schneider-Poetsch. 2002. The evolution of gymnosperms redrawn by phytochrome genes: The Gnetatae appear at the base of the gymnosperms. J. Mol. Evol. **54**:715-724.

Shultz, J. W., and J. C. Regier. 2000. Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade. Proc. R. Soc. Lond. B. Biol. Sci. **267**:1011-1019.

Sidow, A., and W. Thomas. 1994. A molecular evolutionary framework for eukaryotic model organisms. Curr. Biol. **4**:596-603.

Soltis, D. E., P. S. Soltis, and M. Zanis. 2002. Phylogeny of Seed Plants Based on Evidence from Eight Genes. Am. J. Bot. **89**:1670-1681.

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

Stebbins, G. 1974. Flowering Plants: Evolution Above the Species Level. Harv. Univ. Press, Cambridge, MA.

Steel, M., and D. Penny. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. **17**:839-850.

Stefanovic, S., M. Jager, J. Deutsch, J. Broutin, and M. Masselot. 1998. Phylogenetic relationships of conifers inferred from partial 28S rRNA gene sequences. Am. J. Bot. **85**:688-697.

Theissen, G., J. T. Kim, and H. Saedler. 1996. Classification and phylogeny of the MADS-box multigene family suggest defined roles of MADS-box gene subfamilies in the morphological evolution of eukaryotes. J. Mol. Evol. **43**:484-516.

Wettstein, R. R. 1907. Handbuch der Systematischen Botanik. Deuticke, Leipzig, Germany.

Winter, K. U., A. Becker, T. Munster, J. T. Kim, H. Saedler, and G. Theissen. 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.

Zanis, M. J., D. E. Soltis, P. S. Soltis, S. Mathews, and M. J. Donoghue. 2002. The root of the angiosperms revisited. Proc. Natl. Acad. Sci. USA **99**:6848-6853.

Zhang, G., E. Campbell, L. Minakhin, C. Richter, K. Severinov, and S. Darst. 1999. Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 Å resolution. Cell **98**:811-824.

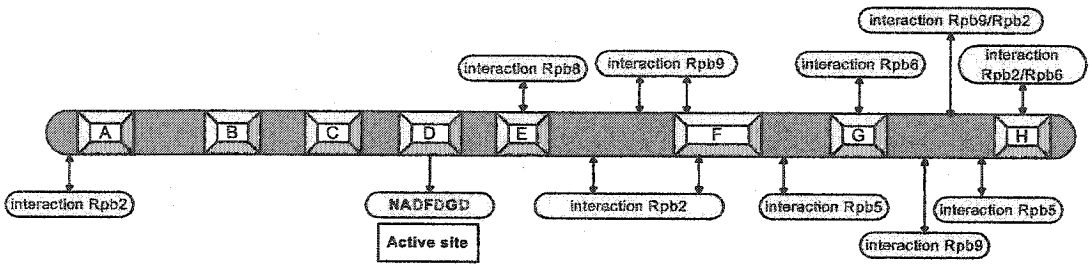
Zillig, W., H. Klenk, P. Palm, G. Puhler, F. Gropp, R. Garret, and H. Leffers. 1989. The phylogenetic relations of DNA-depandent RNA polymerases of archaebacteria, eukaryotes, and eubacteria. Can. J. Microbiol. **35**:73-80.

Table 1. Comparison of the subunits of RNA polymerases

Eukaryotes					
RNAP I	**RNAP II**	**RNAP III**	**Archaea**	**Bacteria**	**Class**
RPA1	RPB1	RPC1	A'+A"	β'	Core
RPA2	RPB2	RPC2	B (B'+B")	β	Core
AC40	RPB3	AC40	D	α	Core
AC19	RPB11	AC19	L	α	Core
RPB6	RPB6	RPB6	K	ω	Core/common
RPB5	RPB5	RPB5	H	-	Common
RPB8	RPB8	RPB8	-	-	Common
RPB10	RPB10	RPB10	N	-	Common
RPB12	RPB12	RPB12	P	-	Common
A12.2	RPB9	C11	X	-	
A14	RPB4	-	F	-	
A43	RPB7	C25	E	-	
+two others	-	+four others	+one other	-	

Note: Core, sequence partially homologous in all RNAPs; common, shared by all eukaryotic RNAPs. The information in this table is retrieved from Table 1 of Cramer (2002).

Figure 1. A schematic diagram of the largest subunit of RNAPs. The active site of the enzyme and the regions of interaction with other subunits are shown on the picture.



Chapter 2

Molecular Biology Methods

Isolation of RNA from plant species

We used RNA as the source of genetic material to ensure amplifying the expressed copies of genes. RNA isolation was done from leaf tissue of all plant species. Three different methodologies were used in the isolation of RNA. These methods include: the Qiagen's RNeasy kit (Mississauga, Ontario); a manual method by Bahloul and Burkard (1993); and a modified protocol developed in this project. This modified protocol (which was also tested in several angiosperm species not used in this work) is presented in detail as a separate section at the end of this chapter.

Amplification of *rpa1* and *rpc1* genes

Designing degenerate primers

I used the Polymerase Chain Reaction (PCR) with degenerate primers to amplify *rpa1* and *rpc1* genes. In this approach, oligonucleotide primers are designed from an alignment of translated amino acid sequences of conserved parts of the target gene in other species.

Due to the degeneracy of the genetic code, certain positions in the translated oligonucleotide are degenerate and the synthetic primer is in fact a mixture of several different molecules that differ in the degenerate positions. The chance of amplifying the desired gene(s) is higher when several genes from closely related organism(s) of interest are available for designing the primers. However, for *rpa1* and *rpc1* there were not many sequences available when I started my project. Another concern was that the conserved blocks of the largest subunit of *rpa1, rpb1* and *rpc1* have many amino acids in common and there was a high chance of amplifying another member of this gene family, *rpb1*, and subsequent problems in identifying and separating the genes. Our goal, therefore, was to design degenerate primers for *rpa1* and *rpc1* in such a way that we could minimize the chance of amplifying *rpb1*. The largest subunit of RNA polymerase I, II and III is about 4 kb long. A ~1.8 kb long fragment from the active site of the enzyme in conserved region D to the conserved region G was selected for PCR amplification (See Figure 1 in Chapter 1). This region was mainly chosen because it was the maximum size we could use to PCR amplify and clone genes. In addition, a smaller region of *rpb1,* covering regions D to F (~1.2 kb), did not provide enough phylogenetic information for plant taxa in another study (M. Sc. thesis, Banoo Malik, 2000. University of Ottawa). I used a universal degenerate forward primer, designed from the amino acid string PYNADFDGDEMN (conserved in most of the RNA polymerases). The nucleotide sequence of this primer is: 5' CCI TAY AAY GCI GAY ITY GAY GGI GAY GAR ATG AA 3'. I designed reverse degenerate primers for region G using a combination of computer and manual methods. The CODEHOP software (Rose et al. 1998) was used for finding candidate oligonucleotides. I used a block of aligned sequences of *rpa1* and *rpc1* as the input. There

were five sequences available in the GenBank for *rpc1* at the time of designing primers (Genbank accession numbers are shown in parenthesis): human (O14802), *Giardia lamblia* (P25202), *Plasmodium falciparum* (A45597), *Trypanosoma brucei* (M27163) and *Saccharomyces cerevisiae* (P04051). As for *rpa1*, five sequences were available in the GenBank: *Drosophila melanogaster* (CAA70321), mouse (O35134), rat (NP_113960), *Trypanosoma brucei* (P16355), *Saccharomyces cerevisiae* (P10964) and *Schizosaccharomyces pombe* (CAA32887). The primers designed by the CODEHOP program were inspected manually and compared with the alignment of conserved block G of *rpa1, rpb1* and *rpc1*. To minimize co-amplifying *rpb1*, I chose three oligonucleotides with a somewhat higher similarity to the *rpa1* and *rpc1* sequences. One of these primers, primer V1 (corresponding to the amino acid string TQMTLKTFHFA), successfully amplified *rpa1* and *rpc1* in PCR reactions (see below) and was used as the reverse primer throughout this project. The sequence of this primer is: 5'GCA AAA TGA AAA GTT TTC AGA GTC ATY TGN GT 3'

Reverse transcriptase polymerase chain reaction (RT-PCR)

Single-stranded cDNA was synthesized using Boehringer Mannheim's 1st Strand cDNA Synthesis Kit following the instructions of the manufacturer. 0.5 to 0.8 μg of RNA was used in each cDNA synthesis reaction for all plant species. Both *rpa1* and *rpc1* were amplified in a single PCR reaction using degenerate primers (see above for the sequences of these primers). PCR conditions were optimized using different combination of

primers. After my initial experiments, I found a reaction consisting of 0.5 μM of primers, 3 mM $MgCl_2$, 2 units of Taq polymerase and 5-10 μl of the cDNA to be optimal for amplifying *rpa1* and *rpc1*. I used a simple strategy to avoid unspecific annealing. The mixture of Taq polymerase and PCR buffer was only added to each sample containing the rest of the reaction mix just before the thermal cycling was started. This approach minimizes the chance of Taq polymerase starting an unspecific annealing. PCR was started by an initial denaturing step (94°C for 5 min) and 35 cycles of denaturing (94°C for 1 min), annealing (45°C or 50°C or 55°C for 1 min) and extension (72°C for 1 min). A final extension of 10 min at 72°C was also performed. PCR products were separated by electrophoresis on agarose gels.

Cloning and sequencing

PCR products of the desired size range were isolated from the gels and were purified using the UltraClean15 kit (MOBIO Laboratories, Carlsbad, CA) according to manufacturer's instructions. Cloning was done using different versions of Invitrogen's TOPO TA Cloning kit (with the PCR 2.1 vector). Table 1 shows that number of clones obtained from each species. Plasmids were purified using Qiagen's Miniprep kit according to manufacturer's instructions. The quantity and quality of plasmids were estimated spectrophotometrically. Sequencing was performed using the Applied Biosystem's Big Dye Terminator Sequencing kit (several versions were used throughout the project) following manufacturer's instructions. Both forward and reverse strands were sequenced using an ABI 310 Genetic Analyzer. Sequences were completed using primer

walking with internal primers designed for each gene (See Appendix 1 for the sequences of internal primers). Sequences were assembled using the software Sequencher (Gene Codes Corporation, Ann Arbor, MI).

An Improved Protocol for the Isolation of RNA from Plants Containing Secondary Metabolites

Pure RNA is important to ensure the successful production of cDNA molecules used in RT-PCR and in making cDNA libraries. Although commercial kits often allow the isolation of RNA from common plant species, more time consuming protocols have to be used with plant species rich in secondary metabolites (Bahloul and Burkard 1993; Lewinshon, Steele, and Croteau 1994; Relle, Sutter, and Wild 1996; Wang, Hunter, and Plant 2000). One of these protocols, developed by Bahloul and Burkard (1993), was shown to perform well with spruce species rich in secondary metabolites. However, this method did not allow us to obtain good quality RNA from the leaves of several plant species from which we wanted to clone RNA polymerase genes using RT-PCR. Here, we report the modifications we made to the Bahloul and Burkard (1993) protocol and show that this improved protocol provides pure RNA from a wide variety of angiosperm and gymnosperm species.

A commercial kit, the RNeasy plant mini kit from Qiagen (Mississuaga, Ontario), allowed us to obtain good quality RNA from the leaves of 7 angiosperm species (*Beta vulgaris*, *Nicotiana tabacum*, *Liriodendron tulipifera*, *Asparagus officinalis*, *Pisum*

sattivum, *Acorus calamus* and *Papaver orientalis*), three gymnosperm species (*Cycas revoluta*, *Ephedra viridis* and *Zamia pumila*) and one fern species (*Psilotum nudum*). Although the Bahloul and Burkard (1993) method allowed us to obtain good quality RNA from the leaves of another 5 angiosperm species (*Persea americanum*, *Saruma henryi*, *Ceratophyllum demersum*, *Drimys winteri* and *Chloranthus oldhami*), we had to modify this method to obtain good RNA from the leaves of the remaining 7 species which contain larger amounts of secondary metabolites (the two angiosperm species *Amborella trichopoda* and *Illicium parviflorum*, and the five gymnosperm species *Taxus canadensis*, *Thuja occidentalis*, *Pinus nigra*, *Gingko biloba*, and *Podocarpus macrophyllus*). These secondary metabolites co-precipitate with RNA and interfere with both RNA isolation and subsequent enzymatic manipulation of these RNA molecules (Bahloul and Burkard 1993; Relle, Sutter, and Wild 1996).

The complete step-by-step protocol is shown in Box 1. All glassware and spatulas were autoclaved (121°C, 15 min.) and baked (200°C for 3hrs) to minimize RNAse contamination. Plant material was frozen in liquid nitrogen and kept at -80°C. Unlike some RNA extraction protocols that use extraction buffer with a relatively high pH (Logemann, Schell, and Willmitzer 1987; Lewinshon, Steele, and Croteau 1994; Relle, Sutter, and Wild 1996; Lönneborg and Jensen 2000; Wang, Hunter, and Plant 2000), the Bahloul and Burkard (1993) protocol uses an extraction buffer with a low pH (5.5). This low pH helps to avoid the ionization and subsequent oxidation of phenolics compounds and to precipitate non-nucleic acid materials (Guillemaut and Marechal-Dourard 1992). It also uses PVP to precipitate polyphenols and cysteine as a safe and efficient anti-oxidant (Bahloul and Burkard 1993).

The main modifications we brought to this protocol is that we replaced the first ethanol precipitation of the RNA by a "pre-precipitation" of the contaminants (polysaccharides and secondary metabolites) using a dilute ethanol solution followed by an isopropanol precipitation of the RNA (steps 8-11) (Lewinshon, Steele, and Croteau 1994). Other modifications include: 1) decreasing the volume of the samples to decrease the amount of interfering secondary metabolites and to make it easier to handle and process them, 2) shortening some steps to decrease RNA degradation and make the protocol more time efficient and 3) adding a DNAse treatment step to ensure that the RNA is not contaminated with DNA.

Total RNA was successfully isolated from multiple samples of all seven species tested. The purity and yield of RNA was measured spectrophotometrically (Table 2). The results were consistent and reproducible. In all cases a single DNase treatment sufficed to remove all DNA contamination (see below). The concentration of RNA in each sample was 200-2000 μg/ml (Table 2). The yield was 33.5-126 μg RNA/g tissue (Table 2), sufficient for more than one hundred RT-PCR reactions. The difference in yields between different species might be due to the difference in nucleic acid content in the leaves of different plant species. We were able to make cDNA using 0.5 to 0.8 μg of RNA in each cDNA synthesis reaction for all of the plants tested. These cDNA molecules allowed us to amplify different RNA polymerase gene regions that were 2-3 kb long (results not shown). This demonstrates that the RNA isolated using this protocol has a high integrity and is not degraded. This RNA was also not contaminated with DNA molecules because PCR amplifications of RNA samples that had not been reverse transcribed did not amplify any products (results not shown). This result was also confirmed by the lack of

introns in the RNA polymerase genes we sequenced. In conclusion, the modified method presented here is simple, relatively fast, and can be used to isolate good quality RNA from a variety of plant species rich in secondary metabolites.

References

Bahloul, M., and G. Burkard. 1993. An improved method for the isolation of total RNA from spruce tissues. Plant Mol. Biol. Rep. **11**:212-215.

Guillemaut, P., and L. Marechal-Dourard. 1992. Isolation of plant DNA: a fast inexpensive and reliable method. Plant Mol. Biol. Rep. **11**:32-37.

Lewinshon, E., C. L. Steele, and R. Croteau. 1994. Simple isolation of functional RNA from woody stems of gymnosperms. Plant Mol. Biol. Rep. **12**:20-25.

Logemann, J., J. Schell, and L. Willmitzer. 1987. Improved method for the isolation of RNA from plant tissues. Anal Biochem. **163**:16-20.

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

Relle, M., A. Sutter, and A. Wild. 1996. A method to isolate cDNA quality RNA from adult conifer needles and a psbA cDNA from Norway Spruce. J. Plant Physiol. **149**:225-228.

Rose, T. M., E. R. Schultz, J. G. Henikoff, S. Pietrokovski, C. M. McCallum, and S. Henikoff. 1998. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. Nuc. Acids Res. **26**:1628-1635.

Wang, S. X., W. Hunter, and A. Plant. 2000. Isolation and Purification of Functional Total RNA from Woody Branches and Needles of Sitka and White Spruce. BioTechniques **28**:292-296.

Box1. RNA isolation protocol

1) Homogenize 0.7 ± 0.20 gram of frozen tissue (-80 °C) vigorously to a fine powder in liquid nitrogen using a mortar and pestle. Plant tissue should be always kept under liquid nitrogen during homogenization. The powdered material is transferred to a 14 ml round bottom sterile disposable tube (or a suitable tube for centrifugation). Let all the liquid nitrogen evaporate. Proceed to step 2 immediately.

2) Add 5 ml of extraction buffer and homogenize by vortexing. Extraction buffer consists of 100 mM sodium acetate, 50 mM EDTA (pH 8.0), 500 mM sodium chloride, 2% PVP w/v (MW 20,000). The pH of this buffer is adjusted to 5.5 and SDS is then added to 1.4% w/v. Solid Cysteine is also added to 10 mM just before use.

3) Incubate in water bath at 65 °C for 10 min with occasional shaking.

4) Centrifuge at 10,000 g at 4 °C for 10 min.

5) Collect the supernatant and add 1/3 volume of 5M potassium acetate (pH 4.8). Mix gently and incubate on ice for 30 min.

6) Centrifuge at 10,000 g at 4 °C for 10 minutes and collect the supernatant.

7) Repeat step 6.

8) Add 1/30 volume 3M sodium acetate and 10% v/v ethanol to the supernatant. Incubate on ice for 10 minutes.

9) Centrifuge at 5,000 g at 4 °C for 20 minutes.

10) Collect the supernatant and add 1/9 v/v 3M sodium acetate and 33% v/v isopropanol. Incubate at -20 °C for 2 hours.

11) Centrifuge at 10,000 g for 30 minutes at 4 °C.

12) Dissolve the pellet in 400 µl water, add 400 µl 4M lithium chloride and precipitate for 2 hours (or overnight) at 4°C.

13) Collect the RNA by centrifugation at 10,000 g at 4°C for 20 min.

14) Wash the pellet twice with 0.5 ml 70% ethanol, air dry for 5 min (or until all ethanol is evaporated), and re-dissolve in 50 µL sterile water.

15) DNAse treat the sample using 2 units of DNAse for at least 20 min (we used the DNA-free kit sold by Ambion Inc., Austin, TX, USA).

Table 1. Sources of plant species, voucher information and number of clones for *rpa1* and *rpc1* genes.

Plant species	Source of specimen; Voucher (greenhouse number)	Number of clones	
		rpa1	rpc1
Amborella trichopoda	U. C. Santa Cruz, CA, USA	4	N/A
Drimys winteri	U Mass Biology Department Greenhouse; Qiu 02090	4	1
Thuja occidentalis	Museum of Nature, Ottawa, Canada; Drouin, Xia, Hajibabaei 2002-2	N/A	2
Cycas revoluta	University of Ottawa Greenhouse; # 35-B	4	5
Ephedra viridis	Carleton University greenhouse; located in greenhouse # 13	4	3
Ginkgo biloba	Sampled from University of Ottawa campus, Osgoode st., Ottawa, Canada	4	2
Nymphaea odorata	Museum of Nature, Ottawa, Canada; Drouin, Malik 1998-1	1	5
Pinus nigra	Museum of Nature, Ottawa, Canada; Drouin, Nickerson 2000-1	4	4
Podocarpus macrophyllus	University of Ottawa Greenhouse; # 111-A	4	3
Psilotum nudum	University of Ottawa Greenhouse; # 90-C	N/A	4
Taxus canadensis	Museum of Nature, Ottawa, Canada; Drouin, Xia, Hajibabaei 2002-3	1	2
Welwitschia mirabilis	Carleton University greenhouse; located in greenhouse # 14	1	4
Zamia muricata	Carleton University greenhouse; located in greenhouse # 13	4	3
Zea mays (Pioneer 3953)	University of Ottawa Greenhouse, Ottawa, Canada	4	N/A

Table 2. Quantity and Quality Measurements of RNA Isolated from Different Plants

Plant group	Plant species	Concentration (μg/ml)	Yield (μg/g)	Absorbance ratio (A_{260}/A_{280})
Angiosperm	Amborella trichopoda	280	56	2.02
Angiosperm	Illicium parviflorum	432.5	43.25	2.00
Gymnosperm	Ginkgo biloba	2050	410	1.95
Gymnosperm	Pinus nigra	200	80	2.01
Gymnosperm	Thuja occidentalis	335	33.5	1.94
Gymnosperm	Taxus canadensis	1260	126	2.00
Gymnosperm	Podocarpus macrophyllus	800	80	1.96

Appendix 1. Sequence information of internal sequencing primers

Plant/Gene	Code	Sequence(5'-3')	Length	Tm	GC%
Amborella/rpa1	AmbAF1	CCAGCCACCATTTACTGTGA	20	59.57	50
Amborella/rpa1	AmbAF2	TGGGCCAACAAGAGTTAGAA	20	58.35	45
Amborella/rpa1	AmbAR1	CAGCCATGCAATGGAAATAA	20	59.52	40
Amborella/rpa1	AmbAR2	CTGCTGCACTTGCACCATAC	20	60.48	55
Cycas/rpa1	CycAF1	GGACGGGAAAACAGGTCATA	20	59	50
Cycas/rpa1	CycAF2	CAAAAGGTGGCTTGGTCAAC	20	65	50
Cycas/rpa1	CycAR1	GACATCGCTGCAAGTAACCA	20	60	50
Cycas/rpa1	CycAR2	CAAATGACCTGCAGCATCC	19	60	52
Cycas/rpc1	CyCF1	GGACAGATTGGAAAGGCAAC	20	59.53	50
Cycas/rpc1	CyCF2	GGCCAAGGGTTTTGTAGCA	19	61.02	52.63
Cycas/rpc1	CyCR1	AGGTCCTCCAAGGCTTTCAT	20	60.07	50
Cycas/rpc1	CyCR2	GAGAGGCGAGTACCAGGTTG	20	59.87	60
Drimys/rpa1	DriAF1	AGACTGGAAGGGATCCCAAG	20	60	55
Drimys/rpa1	DriAF2	TAAGCCCTTTCCGAGAAACT	20	58	45
Drimys/rpa1	DriAR1	GGCCATGCAATGGAAATAAT	20	59	40
Drimys/rpa1	DriAr2	CCATGCATCTGCAAAAACAC	20	60	45
Drimys/rpc1	DriCF05	GATTTGCCAACACCAGCTTT	20	60	45
Drimys/rpc1	DriCF1	GCCGCTGCTTCTTGTATGA	19	60	52
Drimys/rpc1	DriCF2	CCTGCTGCTAAAGGTTTTGT	20	57	45
Drimys/rpc1	DriCf3	CACTCCCGAAGGAGGATG	18	59	61
Drimys/rpc1	DriCR1	CCAGGCCTCATCAATCTACG	20	61	55
Drimys/rpc1	DriCR2	GGCATGTGCCATCATAGTCT	20	58	50
Ephedra/rpa1	EphAF1	GCATGAAGAAACTTGGCAGA	20		
Ephedra/rpa1	EphAf2	GGGCAGCAAGAGTTAGAAGG	20	59	55
Ephedra/rpa1	EphAf3	CTGCTATGGTGAAGACGGAGT	21	59	52
Ephedra/rpa1	EphAR1	CCCACTTCGGGATGTTTTTA	20	60	45
Ephedra/rpa1	EphAr2	CTCTCAATCTCAGCCTGCAT	20	58	50
Ephedra/rpa1	EphAr2	GCCAACTTCTCCCTCCTCTC	20	60	60
Ephedra/rpc1	EphCF1	GGAAATGGAAACAAAAGTGGAC	22	60	41
Ephedra/rpc1	EphCF2	CTGGGTTGACAGCAACTGAA	20	60	50
Ephedra/rpc1	EphCR1	CTTTCATCAATCTACGGGACA	21	58	43
Ephedra/rpc1	EphCR2	CTCAATTTTGCAAGGCGATT	20	60	40
Ginkgo/rpa1	GinAF1	ATGAAGAGGGCTGGTAAGGTC	21	57.56	52.38
Ginkgo/rpa1	GinAF2	GAATCGCGTAACTTCTGAAG	20	52.82	45
Ginkgo/rpa1	GinAR1(N)	TTCGAGATGTTTTGATTGCAG	20	58.91	38.1
Ginkgo/rpa1	GinAR2	TTCCAAACTGTGCCTTGTCA	20	60.28	45
Ginkgorpc1	GinCf1n	GCTGCCTCGTGCATGAAT	18	61	56
Ginkgorpc1	GinCf2n	AATGCAAGTGGTGGGATTGT	20	60	45
Ginkgorpc1	GinCr1n	GTTCAAGGGTGTGCCATCTT	20	60	50
Ginkgorpc1	GinCr2n	AAACACTGCCAGCGACTTCT	20	60	50
Nymphaea/rpa1	NyAF2	TTCTTGCAGATCAGGCACTT	20	60	45
Nymphaea/rpa1	NymAf1.5n1	TGAGAGCATTGGTGAATCTG	20	57	45

Plant/Gene	Code	Sequence(5'-3')	Length	Tm	GC%
Nymphaea/rpa1	NymAf1.5n2	TGAGAGCATTGGTGAATCTGTT	22	60	41
Nymphaea/rpa1	NymAf1n2	TTGACTTCAGAACATCCACACC	22	60	45
Nymphaea/rpa1	NymAf3n1	CAGGAATGGTCACTCTGTGG	20	59	55
Nymphaea/rpa1	NymAf3n1	GGCTCAGGAATGGTCACTCT	20	59	55
Nymphaea/rpa1	NymAr1.5n1	CAGCACGATGGTGCAGTA	18	55	58
Nymphaea/rpa1	NymAr1.5n2	CCAACTCTTGCTGGCCTAAT	20	59	50
Nymphaea/rpa1	NymAr1n1	ATCATGTTGCCCATGGTCTG	20	61	50
Nymphaea/rpa1	NymAr1n2	CATGTTGCCCATGGTCTG	18	60	55
Nymphaea/rpc1	NyCF1	GGCCAGGTCGGGAAAGCTAC	20	63	65
Nymphaea/rpc1	NyCF2	TGCTGCCAAAGGATTTGTAA	20	59	40
Nymphaea/rpc1	NyCR1(N)	CACGCCACTAGCATTACGAA	20	60	50
Nymphaea/rpc1	NyCR2	CGTCATCAATCCCTATCGTG	20	58.95	50
Nymphaearpa1	NymAf1n1	GGCAGCTTGACTTCAGAACA	20	59	50
Pinus/rpa1	PinAr3n	AATGACCTGCTTTCCAGTCC	20	59	50
Pinus/rpa1	PinusAf1n	TCACAAAAGGAAGACCACCT	20	57	45
Pinus/rpa1	PinusAf2n	GATGACGATGACTGGAGCAA	20	60	50
Pinus/rpa1	PinusAr1n	TTGGATAACTGAGCCATCTGAA	22	59	40
Pinus/rpa1	PinusAr2n	AATGAATTTTGTCCCCAAGC	20	59	40
Pinus/rpc1	PinCF1	ACAATGTGCAGCAATGATGG	20	60.55	45
Pinus/rpc1	PinCf2	TGCTTGTGTTGGTCAACAATC	21	59	42
Pinus/rpc1	PinCR1	GCATTGCGTACTGTACCATCA	21	59.6	47
Pinus/rpc1	PinCR2	AAGCCATGATTTCCAATCCA	20	60	40
Podocarpus/rpa1	PodAF1	GGCAAGGTTCCAGGAGATTA	20	59	50
Podocarpus/rpa1	PodAF2	TGCTTGGACAGCAAGAGTTG	20	60	50
Podocarpus/rpa1	PodAF3	TATCCAGCGGCTTGGTAAG	19	59	52
Podocarpus/rpa1	PodAR1	CAAGTACCCACTGCGTGATG	20	60	55
Podocarpus/rpa1	PodAR2	TTCCATCCTCTGCTTGAGGT	20	60	50
Podocarpus/rpa1	PodAR3	CAATTTTGGGGGATGACATT	20	60	40
Podocarpus/rpc1	PodCF1	AAGTTGGGAAGGCCACATT	19	59	47
Podocarpus/rpc1	PodCF2	ACCGCAGTCTTCCACACTTC	20	60	55
Podocarpus/rpc1	PodCR1	GGGTCCATACCATCATCACC	20	60	55
Podocarpus/rpc1	PodCR2	AGTTTGAGCTGCGTTGCAC	19	60	52
Psilotum/rpa1	PsiAf1.5n1	GTGGGGTAGATGACCTGCTA	20	57	55
Psilotum/rpa1	PsiAf1.5n2	TGGGGTAGATGACCTGCTACT	21	59	52
Psilotum/rpa1	PsiAr1.5n1	TGTCCGATTCACCGTTACTTC	21	60	47
Psilotum/rpa1	PsiAr1.5n2	GTAGCCGTGTCCGATTCAC	19	59	58
Psilotum/rpa1	PsilAf1n	TTCCGCTCGAAAAAGAACTC	20	60	45
Psilotum/rpa1	PsilAf1n2	CTGCTATTTGGAAGCCAAGG	20	59	50
Psilotum/rpa1	PsilAf1n3	CCATCCTGCTATTTGGAAGC	20	60	50
Psilotum/rpa1	PsilAf2n	AGAGAGGGGTTGGTGACACA	20	60	55
Psilotum/rpa1	PsilAr1n	GTGTCACCAACCCCTCTCTG	20	60	60
Psilotum/rpa1	PsilAr1n2	CAATTGATCCATCAGCATCC	20	59	45
Psilotum/rpa1	PsilAr1n3	CATCCCGAACCGTGTGAT	18	60	55
Psilotum/rpa1	PsilAr2n	CTTTTTCGAGCGGAAGGATT	20	60	45

Plant/Gene	Code	Sequence(5'-3')	Length	Tm	GC%
Psilotum/rpc1	PsiCF1	CAGCAAAAGTGGTGGTGAAAC	21	58	47.61
Psilotum/rpc1	PsiCF2	CAGCAATCAGTTGGAGGACA	20	59.8	50
Psilotum/rpc1	PsiCR1(N)	GGATCCCTACCATCATCACC	20	59.2	55
Psilotum/rpc1	PsiCR2	AACCCATGATTTCCAATCCA	20	60	40
Taxus/rpa1	TaxAF1	GGAGGATCCACCATTTTTCA	20	60	45
Taxus/rpa1	TaxAf1,5	ACTTGGACAGCAGGAGTTGG	20	60	55
Taxus/rpa1	TaxAf1n	GGGGAGACCGTCTTTTACCA	20	61	55
Taxus/rpa1	TaxAF2	GGGCTTGTTGACACTGCTG	19	60	58
Taxus/rpa1	TaxAf2n	TGTCAACAACAGCCTTTTCC	20	58.75	45
Taxus/rpa1	TaxAR1	GCCCTGCCATACAATGAAAA	20	61	45
Taxus/rpa1	TaxAr1,5	TTTCATAACGTCAACCCCATC	19	60	43
Taxus/rpa1	TaxAr1n	GTCAACCCCATCTTCTCCAT	20	58.7	50
Taxus/rpa1	TaxAR2	GCATCTCTCCCTAACATGTCC	21	59	52
Taxus/rpa1	TaxAr2n	TCATGGTAAAAGACGGTCTCC	21	59	48
Taxus/rpc1	TaxCF1	TTGGAAAAGCGACATTAGGG	20	60	45
Taxus/rpc1	TaxCF2	ACCAGCAGCCAAAGGTTTT	19	60	47
Taxus/rpc1	TaxCR1	CAGGGTCCATACCATCATCC	20	60	55
Taxus/rpc1	TaxCR2	GCCAGGCTGTAGTGTGAGTTT	21	59	52
Taxus/rpc1	TaxCR2	CTGTAGTGTGAGTTTCCCCTTG	22	59	22
Taxus/rpc1	TaxCr2n2	CTGTAGTGTGAGTTTCCCCTTG	22	59	50
Thuja/rpa1	CedAF0.5	GGAAACAGGTCATAACCACAA	21	57	42
Thuja/rpa1	CedAF1	TTCAAACATTGCAAGTCAAAGG	22	60	22
Thuja/rpa1	CedAF2	TGGGTTGCTAAAACCCTTTC	20	59	45
Thuja/rpa1	CedAR1	AAAACGATCCCCAATAAAGC	20	58	40
Thuja/rpa1	CedAR1.5	CAAGTCTGACTGCTGCACAA	20	58	50
Thuja/rpa1	CedAR2	CTTGTCAATCACACCATGCAC	21	60	48
Thuja/rpc1	CedCF1	GCTCGCTAAATTGAGTGCAA	20	59	45
Thuja/rpc1	CedCF2	GCCTGTGTTGGACAACAATC	20	58	50
Thuja/rpc1	CedCF3	TGCTCCCCTGAACCTAGAAA	20	60	50
Thuja/rpc1	CedCR1	TTAACAGGGTCCATGCCATC	20	60	50
Thuja/rpc1	CedCR2	CCTTCTCGTATCTGCTCTTCTT	22	57	45
Thuja/rpc1	CedCR3	TAGGCTTCAAGATGGCAGGT	20	60	50
Welwitschia/rpa1	WelAF2	CTAACTGGCCTTCGCCTTC	19	59.89	57.89
Welwitschia/rpa1	WelAR1	GGAACACGCTTCCCTTCTAA	20	59.31	50
Welwitschia/rpa1	WelAR2	TCCACTGTCTCTTCCCCAAT	20	59.51	50
Welwitschia/rpa1	WelAr2	TCCACTGTCTCTTCCCCAAT	20	59.5	50
Welwitschia/rpc1	WelCF1	CACGATGGATAGGAAATCACG	21	57.97	47.61
Welwitschia/rpc1	WelCF2	GGCGGTTAAAACAGCTGAGA	20	60.39	50
Welwitschia/rpc1	WelCR1(N)	CGCTGTGAGCCCAGAATAA	19	59.96	52.63
Welwitschia/rpc1	WelCR2	AAGTTGCCTTGCCAAGTTGT	20	59.78	45
Zamia/rpa1	ZamAF1	GGGGTGATAGACAAGGCACA	20	60.92	55
Zamia/rpa1	ZamAf2	GTGATGGGCTTGTGGACACT	20	61	55
Zamia/rpa1	ZamAR1	TGATTGCAGTGTCCACAAGC	20	60.88	50
Zamia/rpa1	ZamAr2	CACTGCCTTCAGTTCACCAC	20	59	55

Plant/Gene	Code	Sequence(5'-3')	Length	Tm	GC%
Zamia/rpc1	ZamCF1	GGGAACAAAGATGGGCTTTT	20	60	45
Zamia/rpc1	ZamCF2	CCGTGCACCAAATGGTTT	18	60	50
Zamia/rpc1	ZamCF3	GGCAAAGGCTACTTGTCCTG	20	60	55
Zamia/rpc1	ZamCR1	TCCATGGCATCATCTCCATA	20	60	45
Zamia/rpc1	ZamCR2	AGTGTGAGCTTCCCCTTGTT	20	59	50
Zamia/rpc1	ZamCR3	TTGCATTTGGACGAACAAGA	20	60	40
Zea/rpa1	MaiAF1	AGCTGGGATACTGCTCTTCG	20	59.6	55
Zea/rpa1	MaiAR1(N)	AGAAACCATCCGTGGAACAC	20	59.83	50
Zea/rpa1	MaiAR2	GGCCTTGTCTATTATGCCCTTT	22	60.67	45.45
Zea/rpa1	MazAr2	TCCCATACTTCCCAAACTGG	20	60	50

Chapter 3

Normalized splits: a measure to select and combine DNA data sets for phylogenetic inference

Abstract

We propose a new measure to select genes, or sites of genes, for reconstructing phylogenies. This measure is based on normalized splits values. Normalized splits consist of the number of ways in which the sequences can be grouped into two partitions divided by the length of these sequences. We use six different seed plant genes from nuclear, plastid and mitochondrial genomes, as well as computer simulations, to show that sequences with the lower normalized splits values, given that they contain enough phylogenetic information for a certain divergence level, produce the correct phylogenies.

Introduction

The parallel progress of DNA sequencing, computer science, mathematical biology and molecular evolutionary biology have had a great impact on the advancement of molecular phylogenetics. In fact, the number of publications reporting molecular phylogenies has been increasing exponentially since 1981 (Pagel 1999). On the other hand, it has also become clear that several problems can affect the accuracy of the phylogenies constructed using molecular data (recently reviewed in Sanderson and Shaffer 2002). One of these problems concerns the choice of the genes, or parts thereof, that will produce accurate and statistically supported phylogenetic trees. This is usually performed by choosing orthologous genes that evolve at a rate that is both constant and appropriate to the phylogenetic question being asked (Friedlander, Regier, and Mitter 1992; Graybeal 1994; Yang 1998). Orthologous genes that evolve too slowly will not contain enough phylogenetic information whereas those that evolve too fast will contain saturated sites devoid of phylogenetic signal for all but the closely related sequences found in a given phylogeny. The rate of evolution of third codon (synonymous) positions and first and second codon (nonsynonymous) positions of protein coding genes can also be analyzed separately because synonymous sites usually evolve faster than nonsynonymous sites. It is also important to use genes that evolve at the same rate in different lineages because using genes that evolve at different rates in different lineages might lead to the grouping of unrelated sequences that fortuitously share an accelerated rate of evolution (Kuhner and Felsenstein 1994; Sanderson and Shaffer 2002). It is also clear that many phylogenetic questions cannot be resolved using phylogenies based on a

single gene (e.g., Qiu et al. 1999; Baldauf 2003). It is therefore necessary to develop methods that will allow selecting a variety of phylogenetically informative genes in order to combine their sequences for phylogenetic analyses.

Several methods have been proposed to test the phylogenetic usefulness of different molecular data sets. They include the skewness test statistic (Hillis and Huelsenbeck 1992), spectral analysis (Hendy and Penny 1993; Lockhart, Penny, and Meyer 1995), incongruence length difference (Farris et al. 1995; Dolphin et al. 2000), relative apparent synapomorphy analysis (Lyons-Weiler, Hoelzer, and Tausch 1996), retention index (Naylor and Brown 1997; 1998), likelihood mapping (Strimmer and von Haeseler 1997), quartet mapping (Nieselt-Struwe and von Haeseler 2001), delta plots (Holland et al. 2002) and an entropy-based index of substitution saturation (Xia et al. 2003).

Splits are the number of ways in which the sequences can be grouped into two partitions. For n sequences, the number of bipartitions is given by 2^{n-1} (Hendy and Penny 1993). Lento et al. (1995) pointed out that the ratio of the number of splits found in a set of aligned sequences divided by their length gives an indication of the signal-to-noise ratio for that data set. The data sets having the smallest normalized splits values are expected to have the largest amount of phylogenetic signal relative to the amount of noise. Here we use both known sequences and computer simulations to show that this measure, which we call normalized splits (NS), provides an excellent, yet straightforward, way for exploring the noise in a set of aligned sequences. Normalized splits can therefore be used to select and combine both the genes and the gene partitions that will produce accurate and well-supported phylogenetic trees. However, the concept

of noise is a tricky one when dealing with sequence data. Noise can be stochastic for example due to errors in sequencing, or it can be a result of unforeseen patterns in the process of molecular evolution. For example, maximum parsimony method cannot handle multiple substitutions at one site. Therefore, sites with multiple substitutions could be considered noisy when using the maximum parsimony criterion to reconstruct phylogenies.

The justification for this method is based on a parsimony argument: data sets that contain the smallest proportion of splits should also be those that contain the smallest amount of noise. For example, if all 10 nucleotides of five sequences support the same split, say the split {(1, 2), (3, 4, 5)} that groups these five sequences, the normalized splits value of this data set will be 0.1 (1/10). In contrast, if all 10 nucleotides of five sequences support different splits, the normalized splits value of this data set will be 1 (10/10). Thus normalized splits provide a method to distinguish phylogenetic signal from noise. This method therefore satisfies the requirement of Hillis and Huelsenbeck (1992) in that it allows avoiding the analysis of data that has been randomized with respect to the evolutionary relationships of the sequences being compared.

The sequence data sets we used to test the usefulness of normalized splits consist of six different genes from nuclear, plastid and mitochondrial genomes of eleven plant species (see below). We used these data sets because they are familiar to us and because they have been used to produce a seed plant phylogeny that is currently believed to be accurate (Kuzoff and Gasser 2000). Interestingly, this currently accepted phylogeny was only arrived at when several studies used combined data sets (reviewed in Kuzoff and Gasser 2000). In fact, previous studies based on single genes, such as 18S rRNA and

rbcL genes, produced conflicting phylogenies. For example, the Chase et al. (1993) *rbcL* study supported *Ceratophyllum*, a herbaceous plant, as being the sister group to all other angiosperms whereas the Soltis et al. (1997) 18S rRNA study supported *Amborella*, woody magnoliids and water lilies as being the first-branching angiosperms. It was therefore of interest to us to try to determine why different data sets supported different and incompatible hypotheses for the origin of flowering plants.

Materials and methods

Six genes from eleven plant species were used for this study. The genes used were the *atpB*, *psaA* and *rbcL* plastid genes, the nuclear 18S rDNA genes, the mitochondrial *cox1* genes and the nuclear protein-coding largest subunit of RNA polymerase II (*rpb1*) genes. The plant species consisted of an outgroup (*Psilotum nudum*), four gymnosperm species (*Welwitschia mirabilis*, *Pinus nigra*, *Cycas revoluta* and *Zamia muricata*) and six angiosperm species (*Amborella trichopoda*, *Nymphaea odorata*, *Magnolia soulangeana*, *Oryza sativa*, *Zea mays* and *Arabidopsis thaliana*). The GenBank accession numbers for the *atpB*, *rbcL*, 18S, and *rpb1* genes are listed in Tables 1 and 2 of Nickerson and Drouin (2003). The GenBank accession numbers for the *psaA* and *cox1* genes are listed in Sanderson et al. (2000) and Bowe, Coat, and dePamphilis (2000) respectively. Note that, due to the lack of DNA sequence of *cox1* gene for *Psilotum nudum*, we used the sequence of *Ophioglossum engelmannii* (another fern species) as the outgroup for the *cox1* data set.

Sequences were aligned using ClustalW (Thompson, Higgins, and Gibson 1994) and the DNA sequences of each protein-coding gene was aligned based on their amino

acid sequence alignment. Three data sets corresponding to all three codon positions, first and second codon positions (hereafter referred to as 1+2), and third codon position were built for each protein-coding gene.

Spectral Analysis was performed using the Spectronet software (Huber et al. 2002). Hadamard conjugation (Hendy and Penny 1993) and direct-encoding methods with default settings (as implemented in the Spectronet software) were used to obtain the total number of splits for each data set. Normalized splits were calculated by dividing the total number of splits found in the spectrum divided by the length of the alignment (gaps and missing sites were excluded).

The MEGA2 program was used to obtain the number of synonymous (Ks) and non-synonymous (Ka) substitutions (Kumar et al. 2001). We used the method of Pamilio-Bianchi-Li (Pamilo and Bianchi 1993) to estimate Ks and Ka.

The TreePuzzle program was used for building phylogenetic trees (Strimmer and von Haeseler 1996). Maximum likelihood trees were made using the HKY model (Hasegawa, Kishino, and Yano 1985) with gamma-distributed rates across sites. The gamma shape parameter with eight rate categories, the transition/transversion parameter and nucleotide frequencies were estimated from the data sets.

Simulation was done in two steps. First, a simulated data set was generated using SeqGen (Rambaut and Grassly 1997) following the HKY model with gamma-distributed rates across sites (eight rate categories, gamma shape of 0.5). Second, random noise was added to the simulated data set using a Python script (available upon request). This noise was added only to invariant sites in order to keep the phylogenetic information present in the data sets intact. This approach also allowed us to keep the total number of sites

constant. The script adds random noise to a multiple alignment by randomly selecting conserved sites in the alignment and changing them randomly to one of A, C, G or T nucleotides in each taxa. For example, a site containing nucleotide X in all taxa will be randomly changed to A, C, G or T in each taxon. The noise can be added at a specified percentage of the columns of invariable sites found in a multiple alignment. This approach allows observing the effect that adding an increasing amount of noise has on a given phylogeny. The NS values of these noise-added simulated data sets were only obtained using direct-encoding method. Since our script uses invariant sites to add noise to simulated data sets, and that the Hadamard conjugation method uses these invariant sites in calculating the amount of multiple substitutions (Hendy and Penny 1993; Hendy, Penny, and Steel 1994), we could not use the Hadamard conjugation method in this simulation analysis.

Results

NS values for different genes and partitions

Table 1 shows the NS values of the six genes and, in the case of protein coding genes, their partitions into 1+2 and third bases. Results from both direct-encoding and Hadamard conjugation methods show that stochastic noise varies greatly among the data sets. Although results from the Hadamard conjugation seem to be more sensitive to noise than those of the direct-encoding method, there is a clear agreement between the results of these methods. Among genes, *cox1* and 18S rDNA showed the lowest NS values

(0.065 and 0.094 with the Hadamard conjugation method; 0.043 and 0.058 with the direct-encoding method, respectively) and *rpb1* and *rbcL* showed the highest NS values (0.205 and 0.168 with the Hadamard conjugation method; 0.134 and 0.108 with the direct-encoding method, respectively). Some genes therefore have NS values as much as three times larger than that of other genes.

Among partitions, the most important difference in NS values is seen between the partions of the *rpb1* gene. The 1+2 bases of codons of this gene have NS values that are 5.5 times smaller than those of the third base of codons (Table 1). In fact, the large amount of variation found in the third bases of this gene made it impossible to select the splits from the spectrum of this data set using the Hadamard conjugation method and the NS value for this data set could only be calculated using the direct-encoding method. Other protein coding genes follow a similar pattern. *cox1* has the lowest variation in NS values among the 1+2 and third bases of codons partitions (Table 1).

We examined the correlation of NS values of genes and their length to find out whether NS values are highly influenced by the length of genes. This analysis showed that the NS values of genes, obtained using direct-encoding (Figure 1a) and Hadamard conjugation (Figure 1b), are not strongly correlated with the length of genes. However, the P-values of these correlation analyses were not <0.05. We also tested all the genes and their partitions and found that there is not a strong correlation between the length of sequences and their NS values ($R^2 = 0.085$ and $R^2 = 0.195$ in the direct encoding and the Hadamard conjugation methods, respectively). The P-values of these correlation analyses were not <0.05, however, when the same analyses were done with a larger number of genes and gene partitions (using the data sets that were used in Chapter 4, results not

shown here), the correlations were significant (P=0.02 in both the direct encoding and the Hadamard conjugation methods).

We also studied the correlation of the NS values of genes and their rate of synonymous (Ks) and non-synonymous (Ka) substitutions (Figure 2). We found that NS is correlated with Ks in both the direct-encoding method (Figure 2A) and the Hadamard conjugation method (Figure 2B). The P-values of these correlation analyses were not <0.05 (with the N=5 genes), however when the same analyses were done with a larger number of genes (N=9; using the data sets that were used in Chapter 4), the Ks values were significantly correlated with the NS values (P<0.005 and P<0.05 in the direct encoding method and the Hadamard conjugation method, respectively). However, NS is not correlated with the Ka in either the direct-encoding method (Figure 2A) or the Hadamard conjugation method (Figure 2B)

Combining data sets for phylogenetic analyses

We built phylogenetic trees using the 1+2 bases of codons and the third base of codons of the *rpb1* gene to see if the important difference in NS values of these two data sets would lead to different trees when used in phylogenetic analyses. These two data sets do indeed produce very different trees. The tree obtained with the 1+2 bases of codons of the *rpb1* gene corresponds to the currently accepted phylogeny of the taxa shown in this tree (Figure 3A; Kuzoff and Gasser 2000). All the relationships shown in this tree are also supported by very strong statistical support. In contrast, the tree obtained with the

third base of codons of the *rpb1* gene is not well resolved and well supported and it does correspond to the currently accepted phylogeny of the taxa shown in this tree (Kuzoff and Gasser 2000) (Figure 3B).

We also built phylogenetic trees using the different partitions of the other genes shown in Figure 3 but most of them did not contain enough variable sites to obtain meaningful phylogenies (results not shown). For example, whereas the 1+2 bases of codons and the third base of codons of the *rpb1* gene contain 468 and 927 variable sites respectively, those of the *psaA* gene contain only 190 and 506 variable sites, respectively. We therefore used NS values to combine genes and gene partitions in order to obtain data sets with larger number of variable sites. We then used these data sets for phylogenetic reconstruction. The first combined dataset was composed of 18S, *rbcL* 1+2 positions and *psaA* 1+2 positions (all with NS values smaller than 0.1). This data set contained 595 variable sites. The second combined dataset included third positions of *rbcL*, third positions of *atpB* and third positions of *psaA* (all with NS values greater than 0.198). This data set contained 1061 variable sites. Interestingly, the first combined data set, consisted of genes/partitions with lower NS, produced a tree topology identical to the accepted phylogeny (Figure 3C; Kuzoff and Gasser 2000). However, the tree produced by the second combined data set, consisting of gene partitions with higher NS, did not recover the accepted phylogenetic tree (Figure 3D). This tree supported the grouping of *Nymphaea* and *Amborella* together at the base of angiosperms.

To investigate the amount of phylogenetic signal versus noise that is needed to resolve the phylogeny of examined taxa, we further added the two low NS data sets (used in producing trees in Figure 3A and 3C) and the two high NS data sets (used in producing

trees in Figure 3B and 3D). These two combined data sets contained a total of 1063 (low NS) and 1988 (high NS) variable sites. The low NS data set produced a very strongly supported phylogeny corresponding to the accepted phylogeny (Figure 3E; Kuzoff and Gasser 2000). Interestingly, the high NS data set produced a different topology with much lower support values for several nodes (Figure 3F).

The effect of noise on NS values

In order to complement our analyses based on real data sets, we also generated simulated data sets using the SeqGen program. The strongly supported tree (Figure 3E), obtained using a combination of low NS genes, was used as input tree for the simulation. The same evolutionary model as the one used in TreePuzzle to build the input phylogenetic tree (HKY+ Γ) was used in SeqGen. To mirror the data set used in producing the input tree (Figure 3E), it had a total length of 5860 sites and included 4920 invariable sites (i.e., about 84% of the total sites) and 940 invariable sites. The NS values of the simulated data set, calculated using direct-encoding methods, were 0.024 indicating that, as expected, the simulated data set did not contain a substantial amount of noise (Figure 4). We then added random noise to the simulated data set to evaluate the effect of different levels of noise on the NS values and tree topology. This noise was added only to invariant sites in order to keep the phylogenetic information present in the data sets intact. Given that the simulated data set contained about 84% invariable sites, we therefore were able to add up to 84% noise to them.

Adding stochastic noise to the invariable sites of the simulated data sets clearly increased their NS values and NS values calculated using the direct-encoding method continued to increase by adding more noise (Figure 4).

We built phylogenetic trees using these noise-added simulated data sets (Table 2). Increasing the amount of noise up to 8.4% did not distort the known phylogeny but gradually decreased the support of several nodes (Table 2). Adding more than 33% noise led to incorrect phylogenies with poor support values (Table 2).

Discussion

Both real sequences and computer simulations show that normalized splits provides an excellent, yet straight-forward, way for exploring the stochastic noise in a set of aligned sequences. Normalized splits can therefore be used to select and combine both the genes and the gene partitions that will produce accurate and well-supported phylogenetic trees.

Our results with known sequences show that data sets made up of low NS value sequences can recover the known phylogeny of land plants (Figures 3A, C and E). Furthermore, the support of the nodes of these phylogenies increases as the amount of low NS value sequences increases (Figures 3A, C and E). In contrast, data sets made up of high NS value sequences cannot recover the known phylogeny of land plants (Figures 3B and D) even when a large amount of sequence is used (Figure 3F). Furthermore, the support for most of the nodes of the phylogeny shown in Figure 3F is much lower than the support for the nodes of the phylogeny shown in Figure 3E. This lower support is not

due to the fact that the data set used to construct the tree shown in Figure 3F contains less variable sites than the data set used to construct the tree shown in Figure 3E because they contain 1988 and 1063 variable sites, respectively. The historical signal present in the combined high NS data sets therefore is distorted by noise and a well-supported topology cannot be resolved (Figure 3F). The high support obtained with phylogenetic trees built using data set from genes and partitions with lower noise, and the opposite for the data sets with higher noise, suggests that using NS values to evaluate the noise level of different data sets can give us a valuable guide for combining data sets that will produce strongly supported trees. We can therefore either avoid using noisy data or, when we have no other choice, expect a lower statistical support.

The results obtained with our simulated data sets support our initial hypothesis that NS values increase with the amount of noise added (Figure 4, Table 2). NS values can therefore be used to predict the amount of noise found in a data set. These results also show that whereas relatively small amounts of noise only decrease the support observed for some nodes, adding larger amounts of noise leads to incorrect phylogenies.

There are mixed views about utility of different codon positions in phylogenetics. It is common practice to use only 1+2 codon positions when comparing genes from distantly related species because third codon positions are assumed to be saturated with substitutions and therefore uninformative (Meyer 1994; Swofford et al. 1996; Xia et al. 2003). However, others have argued that third codon positions should be kept because they still contain useful phylogenetic information (Lewis, Mishler, and Vilgalys 1997; Sanderson et al. 2000; Savolainen et al. 2000). Our results suggest that the historical signal in third codon positions, of both nuclear and plastid encoded genes, contains a

large amount of noise. The third codon positions of the *rpb1* genes have a much larger NS value than the 1+2 codon positions (Table 1) and produce an unresolved phylogeny with low statistical support (Figure 3B). We observed the same pattern in the third codon positions of plastid genes. The noise present in the third codon positions of the plastid genes was high enough to produce an imprecise phylogeny with low statistical support (Figure 3D).

Our results also show that the third codon positions of plastid genes *rbcL, atpB* and *psaA* have lower NS values than the third codon position of nuclear *rpb1* (Table 1). Since the synonymous sites of plastid genes are known to evolve roughly four times more slowly than those of nuclear genes (such as the *rpb1* gene), the smaller noise content of third codon position of the plastid genes could be due to their lower synonymous substitution rate (Wolfe et al. 1987, 1989). This is also in agreement with our observation that the NS values are correlated with the rate of synonymous substitutions (Figure 2). The fact that the synonymous sites of mitochondrial genes (such as the *cox1* gene) evolve roughly twelve times more slowly than those of nuclear genes would also explain why the third base of codons of the *cox1* gene have the smallest NS values (Table 1; Wolfe et al. 1987, 1989).

Phylogenetic trees built using maximum parsimony methods often assume that more accurate phylogenies will be obtained using data sets containing the maximum number of parsimony informative sites. Our results show that this is not the case. In all the protein coding genes we analyzed, except the *cox1* gene, the number of parsimony informative sites is three to five time higher in the third base of codons than in the 1+2 bases of codons (Table 1). Yet, the phylogeny obtained with data sets composed of the

third base of the *rpb1* gene or of the *rbcL*, *atpB* and *psaA* genes do not correspond to the known phylogeny of seed plants (Figures 3B and D). Therefore, parsimony informative sites are not necessarily phylogenetically informative sites and NS values provide a straightforward way to select reliable sites for phylogenetic analysis.

Although this study was not designed to address specific issues in seed plant phylogeny, our evaluation of the noise level in different plant genes is relevant to plant molecular phylogenies. Some of the topological discrepancies and low statistical supports leading to different hypotheses in recent plant phylogenetic studies might be a result of using combined data sets of genes with different noise levels. The common practice of adding more genes and taxa for building molecular phylogenies, reflected in a number of recent studies (e.g., Qiu et al. 1999), has helped in resolving a number of phylogenetic problems. It also produced a number of new hypotheses with no apparent conclusion (Kuzoff and Gasser 2000). Our study suggests that simply adding more genes and taxa may not lead to a well-supported phylogeny especially if the amount of noise is substantial enough to mask or distort the phylogenetic signal. For example, the topology produced using a combined data set with high noise level shows a clade containing *Amborella* and *Nymphaea* as the sister group to the other angiosperms (Figure 3D). In contrast, the topology produced using data sets with low noise levels (Figures 3A, C and E) agrees with the currently accepted phylogeny which considers *Amborella* to be the only sister group to all other angiosperms (Kuzoff and Gasser 2000).

Acknowledgments

We would like to thank Greg A. Singer for his help with our Python script.

References

Baldauf, S. L. 2003. The deep roots of eukaryotes. Science **300**:1703-1706.

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

Chase, M., D. Soltis, R. Olmstead, D. Morgan, D. Les, M. Duvall, R. Price, H. Hills, Y. Qiu, K. Kron, J. Rettig, E. Conti, J. Palmer, M. Clegg, J. Manhart, K. Sytsma, H. Michaels, W. Kress, M. Donoghue, W. Clark, M. Hedren, B. Gaut, R. Jansen, K.-J. Kim, C. Wimpee, J. Smith, G. Furnier, S. Straus, Q. Xiang, G. Plunkett, P. Soltis, L. Eguiarte, G. Learn, S. Barrett, S. Graham, and V. Albert. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. Ann. Mo. Bot. Gard. **80**:528-580.

Dolphin, K., R. Belshaw, C. Orme, and D. Quicke. 2000. Noise and incongruence: Interpreting results of the incongruence length difference test. Mol. Phylogenet. Evol. **17**:401-406.

Farris, J. S., M. Källersjö, A. G. Kluge, and C. Bult. 1995. Testing significance of incongruence. Cladistics **10**:315-319.

Friedlander, T. P., J. C. Regier, and C. Mitter. 1992. Nuclear gene sequences for higher level phylogenetic analysis: 14 promising candidates. Syst. Biol. **41**:483-490.

Graybeal, A. 1994. Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. Syst. Biol. **43**:174-193.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **21**:160-174.

Hendy, M. D., and D. Penny. 1993. Spectral analysis of phylogenetic data. J. Classification **10**:5-24.

Hillis, D. M., and J. P. Huelsenbeck. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. J. Hered. **83**:189-195.

Holland, B., K. Huber, A. Dress, and V. Moulton. 2002. Delta plots: a tool for analyzing phylogenetic distance data. Mol. Biol. Evol. **19**:2051-2059.

Huber, K. T., M. Langton, D. Penny, V. Moulton, and M. D. Hendy. 2002. Spectronet: A package for computing spectra and median networks. Applied Bioinformatics **1**:159-161.

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

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17**:1244-1245.

Kuzoff, R. K., and C. S. Gasser. 2000. Recent progress in reconstructing angiosperm phylogeny. Trends Plant Sci. **5**:330-336.

Lento, G. M., R. E. Hickson, G. K. Chambers, and D. Penny. 1995. Use of spectral analysis to test hypotheses on the origin of pinnipeds. Mol. Biol. Evol. **12**:28-52.

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

Lockhart, P. J., D. Penny, and A. Meyer. 1995. Testing the phylogeny of swordtail fishes using split decomposition and spectral analysis. J. Mol. Evol. **41**:666-674.

Lyons-Weiler, J., G. A. Hoelzer, and R. J. Tausch. 1996. Relative apparent synapomorphy analysis (RASA). I: The statistical measurement of phylogenetic signal. Mol. Biol. Evol. **13**:749-757.

Meyer, A. 1994. Shortcomings of the cytochrome b gene as a molecular maker. TREE **9**:278-280.

Naylor, G. J., and W. M. Brown. 1997. Structural biology and phylogenetic estimation. Nature **388**:527-528.

Naylor, G. J. P., and W. M. Brown. 1998. Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparison of sequences. Syst. Biol. **47**:61-76.

Nickerson, J., and G. Drouin. 2003. The sequence of the largest subunit of RNA polymerase II is a useful maker for inferring seed plant phylogeny. Mol. Phylogenet. Evol., in press

Nieselt-Struwe, K., and A. von Haeseler. 2001. Quartet-mapping, a generalization of the likelihood-mapping procedure. Mol. Biol. Evol. **18**:1204-1219.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature **401**:877-884.

Pamilo, P., and N. Bianchi. 1993. Evolution of the Zfx and Zfy genes : rates and interdependence between the genes. Mol. Biol. Evol. **10**:271-281.

Qiu, Y. L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. 1999. The earliest

angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature **402**:404-407.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13**:235-238.

Sanderson, M. J., and H. B. Shaffer. 2002. Troubleshooting molecular phylogenetic analyses. Annu. Rev. Ecol. Syst. **33**:49-72.

Sanderson, M. J., M. F. Wojciechowski, J. M. Hu, T. S. Khan, and S. G. Brady. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol. Biol. Evol. **17**:782-797.

Savolainen, V., M. W. Chase, S. B. Hoot, C. M. Morton, D. E. Soltis, C. Bayer, M. F. Fay, A. Y. de Bruijn, S. Sullivan, and Y. L. Qiu. 2000. Phylogenetics of flowering plants based on combined analysis of plastid *atpB* and *rbcL* gene sequences. Syst. Biol. **49**:306-362.

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

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

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

Swofford, D. L., G. L. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407-514 *in* D. M. Hillis, C. Morowitz, and M. B.K., eds. Molecular Systematics. Sinauer Associates, Sunderland, Mass.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 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.

Wolfe, K.H., W.-H. Li, and P.M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. Proc. Natl. Acad. Sci. USA **84**: 9054-9058.

Wolfe, K.H., P.M. Sharp, and W.-H. Li. 1989. Rates of synonymous substitution in plant nuclear genes. J. Mol. Evol. **29**: 208-211.

Xia, X., Z. Xie, M. Salemi, L. Chen, and Y. Wang. 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol. **26**:1-7.

Yang, Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. **47**:125-133.

Table 1. Normalized splits values for six genes and their partitions

Data set	Sites	Variable sites	Parsimony informative sites	Splits (D)	NS (D)	Splits (H)	NS (H)
18S	1696	325	177	99	0.058	160	0.094
rpb1	3022	1395	1069	406	0.134	619	0.205
rpb1 (1+2)	2016	468	256	138	0.068	192	0.095
rpb1 (third)	1006	927	813	377	0.374	N/A	N/A
rbcL	1074	345	225	116	0.108	180	0.168
rbcL (1+2)	716	80	41	42	0.059	49	0.068
rbcL (third)	358	265	184	99	0.222	445	1.243
atpB	1161	376	240	108	0.093	197	0.170
atpB (1+2)	774	86	40	33	0.043	40	0.052
atpB (third)	387	290	200	100	0.258	379	0.979
psaA	2143	696	443	155	0.072	244	0.114
psaA (1+2)	1428	190	103	60	0.042	71	0.050
psaA (third)	715	506	340	142	0.199	289	0.404
cox1	1404	409	196	61	0.043	84	0.060
cox1 (1+2)	936	180	100	41	0.044	53	0.056
cox1 (third)	468	229	96	42	0.090	56	0.120

Notes. Sites, total number of sites; Variable sites, number of variable sites; Parsimony informative sites, number of parsimony informative sites; NS, normalized splits; (D) denotes direct-encoding method; (H) denotes Hadamard conjugation; N/A denotes that calculation was not possible using Hadamard conjugation. First and second positions of codons are denoted by (1+2) and the third positions of codons are denoted by (third).

Table 2. The effect of adding noise on phylogenies reconstructed from simulated data

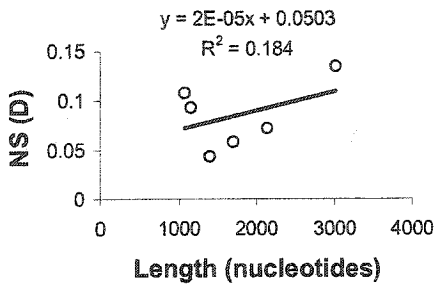
% Noise added	No. of variable sites	NS (D)	Reconstructed Tree
0.0	940	0.024	(Ps,((Cy,Za)100,(Pi,We)100)100,(((((Ze,Or)100,Ar)100,Ma)100,Ny)97,Am)100);
2.1	1064	0.038	(Ps,(((((Ze,Or)100,Ar)100,Ma)100,Ny)99,Am)100,((Cy,Za)100,(Pi,We)100)97);
4.2	1187	0.051	(Ps,(((((Ze,Or)100,Ar)100,Ma)100,Ny)91,Am)100,((Cy,Za)100,(Pi,We)100)99);
8.4	1433	0.073	(Ps,(((((Ze,Or)100,Ar)99,Ma)100,Ny)99,Am)100,((Cy,Za)100,(Pi,We)98)95);
16.8	1925	0.103	(Ps,(Cy,Za)100,((((((Ze,Or)100,Ar)96,Ma)99,Ny)95,Am)100,(Pi,We)93);
25.2	2417	0.124	(Ps,(((((Ze,Or)100,Ar)100,Ma)98,Ny)95,Am)100,((Cy,Za)100,(Pi,We)100)54);
33.5	2909	0.136	(Ps,(Pi,We)99,((((((Ze,Or)100,Ar)75,Ma)97,Ny)80,Am)99,(Cy,Za)97)86);
42.0	3401	0.149	(Ps,((((Ze,Or)100,Ar)75,Ma)99,(Ny,Am)83)100,((Cy,Za)99,(Pi,We)66)55);
50.4	3893	0.156	(Ps,((((Ze,Or)100,Ar)73,Ma)95,Ny,Am)100,((Pi,We)93,(Cy,Za)89)87);
58.8	4385	0.160	(Ps,((((Ze,Or)100,Ar)86,(Ma,Ny)53)80,Am)98,((Cy,Za)92,(Pi,We)88)80);
67.2	4877	0.164	(Ps,(Cy,Za)97,(((((Ze,Or)100,Ar)80,Ma)99,Ny)56,Am)99,(Pi,We)85)68);
75.5	5369	0.167	(Ps,(((Ze,Or)70,Ar,Ma,Ny)78,Am)94,(Pi,We)91,(Cy,Za)77);
83.9	5860	0.169	(Ps,((((Ze,Or)100,Ar)94,Ma)80,(Ny,Am)58)94,(Pi,We)92,(Cy,Za)90);

Notes. names of taxa from the original tree, used for generating simulated data sets in Seq-Gen, are represented in the nodes of the above trees. The abbreviations for name of taxa are: Am: Amborella; Ar: Arabidopsis; Cy: Cycas; Ma: Magnolia; Ny: Nymphaea; Or: Oryze; Pi: Pinus; Ps: Psilotum; We: Welwitschia; Za: Zamia; Ze: Zea

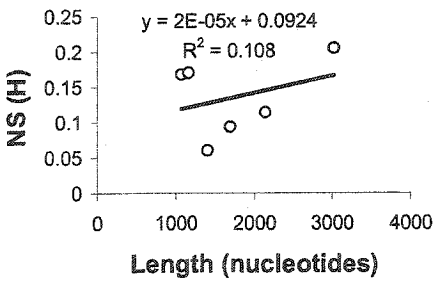
NS (D) denotes normalized splits using direct-encoding method. Shaded parts of the table represent trees with wrong or unresolved topologies. The data sets contained a total of 5860 sites and the initial number of variable and invariable sites were 940 and 4920 respectively.

Figure 1. Correlation of NS and the length of sequences.

A



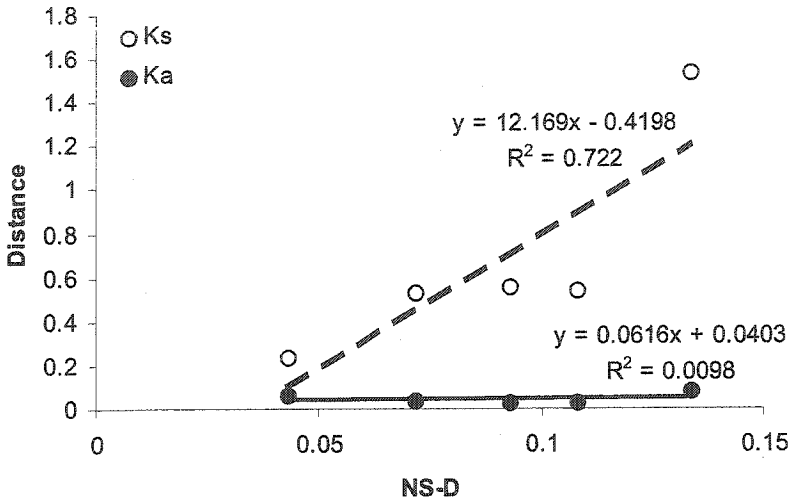
B



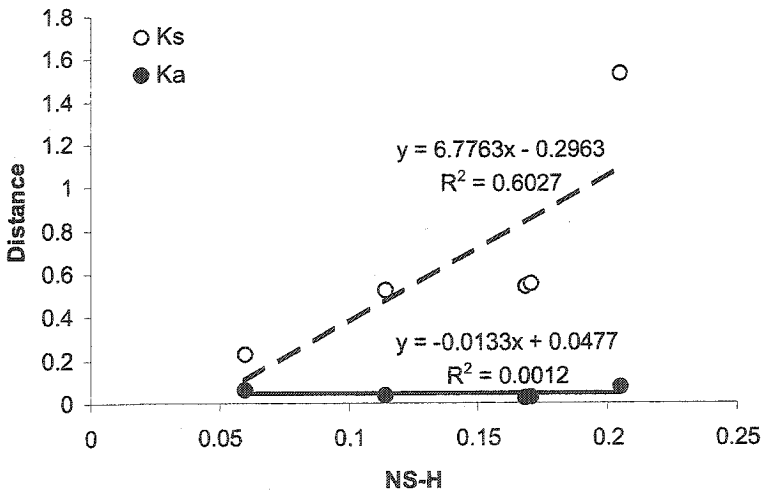
Notes. (D) denotes direct-encoding method; (H) denotes Hadamard conjugation method.

Figure 2. Correlation of the NS values of genes with their number of synonymous (Ks) and non-synonymous (Ka) substitutions.

A



B



Note. (D) denotes direct-encoding method; (H) denotes Hadamard conjugation method.

Figure 3. Phylogenetic trees reconstructed using data sets composed of genes/partitions with low (a, c and e) and high (b, d and f) NS values. a, tree based on *rpb1* 1+2 codon positions; b, tree based on *rpb1* third codon positions; c, tree based the combination of 18S, *rbcL* 1+2 codon positions and *psaA* 1+2 codon positions; d, tree based on the combination of *rbcL* third codon positions, *atpB* third codon positions and *psaA* third codon positions; e, tree based the combination of data sets used in (a) and (c); f, tree based the combination of data sets used in (b) and (d). The scale bars represent 10% substitutions. Trees are made using the TreePuzzle program. Statistical supports from the TreePuzzle are shown on branches.

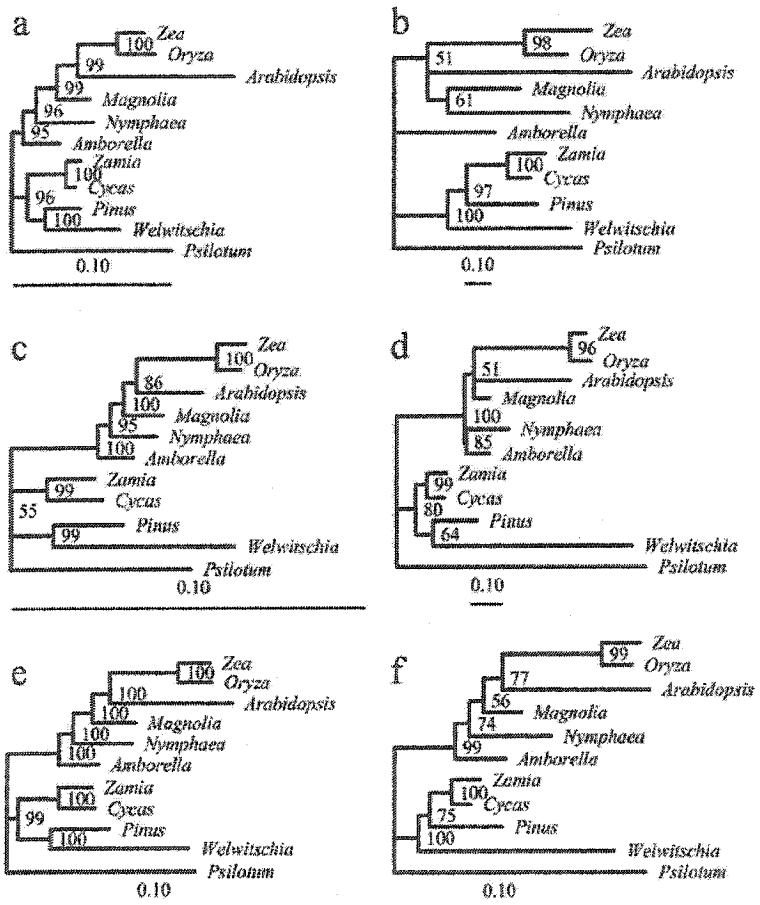
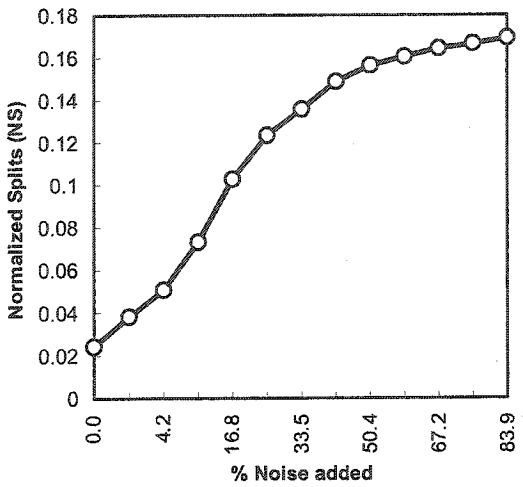


Figure 4. Effect of adding noise on NS values of simulated sequences. NS is calculated using direct-encoding method.



Chapter 4

Are Gnetales derived from conifers? Evidence from three nuclear RNA polymerases and seven other genes using spectral analysis and phylogenetic reconstructions

Abstract

Seed plant phylogeny and in particular the position of Gnetales have been problematic and controversial for a long time. We used the sequence information from three nuclear encoded RNA polymerases together with seven other genes from chloroplast, mitochondria and 18S rRNA (16223 nucleotides) to reconstruct seed plant phylogeny and to evaluate the recently proposed Gnepines hypothesis. Our results, with normalized splits and evolutionary rates, showed that the third codon positions of nuclear RNA polymerases and chloroplast genes can cause erroneous phylogenies. Spectral analysis supported the Gnepines groupings in the spectrum of all protein coding genes. 18s rRNA, however, did not support this hypothesis. Our phylogenetic reconstructions generally supported the Gnepines hypothesis. Our maximum likelihood analysis supported the Gnepines hypothesis with high bootstrap values. However, the use of third codon positions of RNA polymerase genes and chloroplast genes especially in maximum

parsimony analysis resulted in various contradictory phylogenies. By removing these data partitions, we reached a well-supported tree in maximum parsimony analysis, in which the Gnepines hypothesis is strongly supported. Our work clarifies the position of Gnetales, as a sister group to Pinaceae, as well other branches of seed plant phylogeny, and provides insight for addressing complex phylogenies using combined data sets from different genomic compartments.

Introduction

Evolutionary relationships of seed plants have been a controversial and problematic issue for several decades. Seed plants radiated rapidly and long ago. The five living groups of seed plants (angiosperms, conifers, cycads, Gnetales and Ginkgo) diverged several hundred million years ago and evolved independently since then (Kenrick and Crane 1997). Therefore, tracing the phylogeny of these groups is difficult and requires homology assessment of a large character sampling. One of the long standing controversies in plant phylogny is the position of Gnetales, the least familiar group of seed plants. Although parsimony based studies of all available morphological characters clarified the outline of major land plant groupings (e.g., Crane 1985; Doyle and Donoghue 1992), they were unable to reach a consensus on a detailed phylogeny. A major conclusion from morphological studies, however, was support for the Anthophyte hypothesis that grouped Gnetales with angiosperms (Crane 1985; Doyle and Donoghue 1986; Doyle and Donoghue 1992). On the other hand, molecular studies based on chloroplastic *rbcL* genes (Chase et al. 1993), and 18S rRNA nuclear ribosomal gene (Hamby and Zimmer 1992; Doyle, Donoghue, and Zimmer 1994), 28s rRNA nuclear ribosomal gene (Stefanovic et al. 1998) and chloroplast ITS sequences (Goremykin et al.

1996) raised serious doubts about the Anthophyte hypothesis. However, their results were inconclusive because they had low statistical support and lacked congruence among different genes and methods of analysis (Donoghue and Doyle 2000). Most of these analyses were based on parsimony criterion, which is known to falsely group long branches of a tree (Felsenstein 1978). In fact some of the above mentioned studies reported different groupings using different taxa as outgroups (Doyle, Donoghue, and Zimmer 1994).

With the advancement of phylogenetic methods and sequencing technology, several new molecular studies have offered insights in seed plant phylogeny and in particular the position of Gnetales. However, the issue remains unresolved and controversial. A major conclusion of these new molecular phylogenetic analyses was the rejection of the Anthophyte hypothesis (Hansen et al. 1999; Winter et al. 1999) and grouping the Gnetales with conifers. The most recent hypothesis on the position of Gnetales was offered, exclusively based on molecular markers, by Chaw et al. (2000) and supported by Bowe, Coat, and dePamphilis (2000). This hypothesis, which is called the Gnepines hypothesis, places Gnetales within conifers as a sister group to Pinaceae. Chaw et al. (2000) used mitochondrial small subunit rRNA genes (mtSSU rRNA) and nuclear small subunit rRNA genes (nuSSU rRNA) in combination with *rbcL* for all groups of seed plants. Their mtSSU rRNA dataset supported the grouping of Gnetales and Pinacea in maximum likelihood and maximum parsimony methods with high bootstrap support (98 and 97 respectively). Bowe, Coat, and dePamphilis (2000) also used two mitochondrial genes, *cox1* and *atpA*, together with known sequences of *rbcL* and 18S rRNA in a data set consisted of all groups of seed plants. Their maximum likelihood,

maximum parsimony and neighbor joining trees supported the Gnepines hypothesis. The Gnepines hypothesis, which raised a lot of criticism in the plant systematics community (Donoghue and Doyle 2000), has been tested and challenged by several phylogenetic studies (Sanderson et al. 2000; Rydin, Källersjö, and Friis 2002; Schmidt and Schneider-Poetsch 2002; Soltis, Soltis, and Zanis 2002). None of these studies, however, could resolve the position of Gnetales. Rydin, Källersjö, and Friis's (2002) study, which rejected the Gnepines hypothesis, was based on a parsimony analysis of *rbcL*, *atpB*, 18S and 26S rRNA genes. This work is interesting since it proposes that third codon positions of chloroplast genes contain most of the phylogenetic information and should therefore be used for phylogenetic reconstructions. They also pointed out that there is a conflict between transitions and transversions within each codon position. Rydin, Källersjö, and Friis (2002) however, do not reach a definitive conclusion about the position of Gnetales. Their trees reject the Gnepines hypothesis and places Gnetales either as the sister group to all other seed plants or angiosperms (based on the position of the root). In contrast to Rydin, Källersjö, and Friis (2002), Soltis, Soltis, and Zanis (2002) generally support the Gnepines hypothesis. They used eight genes in ML, MP and Bayesian phylogenetic methods. They obtained different topologies depending on the choice of data set or the method of analysis. Interestingly, this study suggests that most of the discrepancies in seed plant phylogenies involve third codon positions of chloroplast genes (their data set did not include any nuclear protein coding gene). In an analysis of nuclear encoded phytochrome genes, Schmidt and Schneider-Poetsch (2002), rejected the Gnepines hypothesis, but could not reach a conclusion on the position of Gnetales. Magallón and Sanderson's (2002) work used the first and second codon positions of the highly

conserved chloroplast *psaA* and *psaB* genes. Their data supported the Gnepines hypothesis in maximum likelihood and maximum parsimony analyses. However, data sets of third codon positions and all three codon positions, although inconclusive on the position of Gnetales, rejected the Gnepines hypothesis.

Spectral analysis (Hendy and Penny 1993) offers a unique possibility to evaluate genes/data partitions (Chapter 3) and to test phylogenetic hypotheses (Lento et al. 1995) without building a phylogenetic tree. This method is based on splits which are the number of ways in which the sequences can be grouped into two partitions. We recently showed, using both real and simulated sequences, that the data sets having the smallest normalized splits (NS) values are expected to have the lowest amount of (random) noise (Chapter 3). Therefore these low NS data sets can be used in producing more accurate phylogenies. Spectral analysis can also help in assessing support for particular phylogenetic groupings in a given data set (Lento et al. 1995; Lockhart, Penny, and Meyer 1995). This is achieved by measuring the observed or corrected frequency of support as well as contradictory signals for a particular split in a data set. In this paper we use spectral analysis both in evaluating data sets and data partitions and in showing supporting/contradictory signals for three groupings regarding the relative position of Gnetales to conifers. Figure 1 shows these groupings in a phylogenetic tree as well as a split system.

To be able to make a large data set from protein coding genes of all three genomic compartments, we cloned and sequenced the genes encoding the largest subunit of three nuclear encoded RNA polymerases (*rpa1, rpb1, rpc1)* and used them in combination with chloroplast, mitochondrial and 18S rRNA genes. *rpa1, rpb1* and *rpc1* are the central part

of the transcription machinery in eukaryotic cells (the active site of the RNA polymerase enzyme is located in the largest subunit). These genes belong to the RNA polymerase gene family and are conserved over a wide evolutionary scale (from *E. coli* to human). However, they evolve with different rates (Iwabe et al. 1991; see Chapter 5) and therefore can provide resolution for different divergence levels. *rpb1* has been used in resolving a number of phylogenetic problems (e.g., Stiller and Hall 1997; Hirt et al. 1999; Stiller, Riley, and Hall 2001) and was recently shown to be a useful marker for plant phylogeny (Nickerson and Drouin in press). In this report we introduce *rpa1* and *rpc1* as new molecular phylogenetic markers.

This work offers insight from new genes and methods of analyses to the old problem of seed plant phylogeny and in particular the phylogenetic position of Gnetales. We show that third codon positions in RNA polymerases and chloroplast genes are affected by noise and can produce erroneous phylogenetic trees. We further show, using spectral analysis, that there is support for the Gnepines hypothesis in different genes from nucleus, chloroplast and mitochondria. Our phylogenetic reconstructions further supports the position of Gnetales as nested within conifers and as a sister group to Pinaceae.

Materials and methods

This study used nine protein coding genes (three nuclear, four plastid and two mitochondrial) and nuclear ribosomal gene 18S rRNA (total sequence length of 16223 nucleotides) from all major groups of seed plants (twelve taxa). The sequences of chloroplast *psaA*, *psbB*, *rbcL* and *atpB* genes, mitochondrial *cox1* and *atpA* genes, nuclear 18S rRNA genes and seven *rpb1* sequences were retrieved from GenBank. Plants species

and Genbank accession numbers, as well as taxon substitutions to allow combined multigene analyses, are listed in Appendix 1. To ensure sequencing functional copies of nuclear RNA polymerase genes, total RNA was extracted for all twelve plant taxa using either QiaGen's Rneasy or a manual method (See Chapter 2). *rpb1* genes for five taxa were obtained using a published protocol (Nikerson and Drouin 2003). *rpa1,* and *rpc1* genes were amplified using degenerate primers (for sequences of primers and PCR conditions see Appendix 2). We used Invitrogen's TOPO TA cloning kit to clone each gene. Sequencing was performed for both forward and reverse strands using an ABI 310 genetic analyzer. Sequences were assembled using the Seqencher program (Gene Codes Corporation, Ann Arbor, MI).

Sequences were aligned using ClustalW (Thompson, Higgins, and Gibson 1994) and the DNA sequences of protein-coding genes were aligned based on their amino acid sequence alignment. Three data sets corresponding to all three codon positions, first and second codon positions (hereafter referred to as 1+2), and third codon position were built for each protein-coding gene.

Spectral Analysis was performed using the Spectronet software (Huber et al. 2002). Direct-encoding and Hadamard conjugation (Hendy and Penny 1993; Hendy, Penny, and Steel 1994) as implemented in the Spectronet software, were used to obtain the total number of splits for each data set. NS values were calculated by dividing the total number of splits found in the spectrum by the length of the alignment.

To study the signals for the position of Gnetales relative to conifers (see Figure 1) we produced the spectrum of each gene for a set of aligned sequences of all gymnosperms (9 taxa) using direct-encoding (observed) as well as Hadamard conjugation

(which makes corrections for unobserved multiple substitutions). The outgroup and angiosperms were not included in this analysis to reduce the complexity of the spectrum. The numerical value for the frequency of support was calculated (as implemented in the Spectronet software) by dividing the sum of occurrences for a split by the total number of sites in the alignment (Lento et al. 1995). The conflict value for each split is the sum of all other splits that contradict the partitioning of taxa in the first split. Since a split may be contradictory to many splits, the conflict value may be much larger than the support. We followed Lento's normalization method for the frequency of conflict (Lento et al. 1995). This method divides the sum of all support values by the sum of all conflict values. Each conflict value is then multiplied by this ratio.

Evolutionary distances were calculated using the MEGA2 program (Kumar et al. 2001). We used the method of Pamilo-Bianchi- Li (Pamilo and Bianchi 1993) to calculate the number of synonymous and non-synonymous substitutions. We used K2P model (Kimura 1980) and a gamma distribution to calculate the number of pairwise (between individual genes, for example *rpb1* of *Ginkgo* and *rpb1* of *Psilotum*) nucleotide substitutions for 1+2 and third codon positions.

Maximum likelihood (ML) and maximum parsimony (MP) phylogenetic trees (exhaustive searches) were reconstructed using PAUP* Version 4 (Swofford 2000). MODELTEST (Posada and Crandall 1998) was used to obtain the model parameters for likelihood analysis for each data set. In all data sets the general time reversible model was selected by the MODELTEST as the best model for the ML analyses. Statistical support for each topology was obtained using full heuristic bootstrapping (100 replicates for ML and 1000 replicates for MP) as implemented in PAUP*. Several combinations of the data

sets were used to obtain phylogenetic trees using both ML and MP. These combinations are described in the results section.

Results

Spectral analysis and rates of molecular evolution

Table 1 summarizes the NS values for each gene and gene partitions into 1+2 codon positions and third codon position. Among genes nuclear *rpc1* has the highest NS (0.147 and 0.245 in the direct method and the Hadamard conjugation method, respectively) and mitochondrial *atp*A has the lowest NS (0.080 and 0.109 in the direct method and the Hadamard conjugation method, respectively). Among data partitions, third codon positions of nuclear RNA polymerases and chloroplast genes show dramatically higher NS values than 1+2 codon positions of these genes (it was not possible to calculate the splits for the third codon positions of RNA polymerase genes using Hadamard conjugation). The difference between NS values of 1+2 and third codon positions of mitochondrial genes is much lower. Table 1 also shows the number of splits from each data partition as well as the number of variable and parsimony informative sites. Nuclear RNA polymerase data set include 3133 variable sites and 2104 parsimony-informative sites. Among these, 1227 variable and 621 parsimony-informative sites belong to the 1+2 codon position and 1906 variable and 1483 parsimony-informative sites belong to third codon positions.

We also calculated the number of substitutions in different genes and data partitions. Table 2 summarizes these calculations for the number of synonymous and non-synonymous substitutions for each protein-coding gene and combined data sets of

RNA polymerases, chloroplast and mitochondrial genes. Among genes, nuclear *rpb1* showed the highest (1.223±0.036) and mitochondrial *cox1* showed the lowest (0.341±0.018) number of synonymous substitutions. Nuclear *rpa1* showed the highest (0.189±0.012) and chloroplast *rbcL* showed the lowest (0.020±0.003) number of non-synonymous substitutions. Among combined data set from different genomic compartments, nuclear RNA polymerases showed a higher number of synonymous substitutions (1.188±0.025), followed by chloroplast (0.502±0.012) and mitochondrial (0.341±0.014) genes. For non-synonymous substitutions, nuclear RNA polymerases showed the highest value (0.097±0.003), followed by mitochondrial genes (0.081±0.005) and chloroplast genes (0.031±0.002). We also calculated pairwise distances in 1+2 codon positions as well third codon positions of nuclear RNA polymerases, chloroplast and mitochondrial genes (Figure 2). Interestingly, the majority of the distances in the third codon position of nuclear RNA polymerases and chloroplast genes are close to or above 1 substitution per site (Figure 2A). This is not the case in the third codon position of mitochondrial genes where only a small fraction of the distances are above 1 substitution per site (Figure 2A). Interestingly, in nuclear RNA polymerases and chloroplast genes, the majority of the distances between pairs of sequences where Gnetales species are one of the pairs are among the higher end of the graphs. These data points are shown in black on Figure 2A. The 1+2 codon positions in RNA polymerases, chloroplast and mitochondrial data sets, however, all have distance values below 0.25 substitution per site (Figure 2B). However, distances involving Gnetales species are in the higher end of graphs.

Figure 3, 4 and 5 show the spectra of *rpa1*, *rpb1* and *rpc1* genes for an alignment of the 9 gymnosperm taxa we used. We also made the spectra of the other genes, however, only the full spectrum of two other genes, *psaA* and 18S rRNA (Figure 6 and 7 respectively) are shown for comparison with *rpa1*, *rpb1* and *rpc1*. Signal/conflict frequencies supporting three groupings regarding the relative position of Gnetales and conifers (A, B and C on Figure 1) are color-coded (shaded) on the spectra. Other signals are also shown on the spectrum of each gene (all with the same color). In each spectrum, there are 8 signals that do not have conflict values. These signals correspond to the singleton splits (separating one taxa from the others). Direct-encoding (representing observed values) and Hadamard-corrected spectra are shown in two panels. The Spectrum of *rpa1*, *rpb1* and *rpc1* generally show the same signal/conflict pattern for the placement of Gnetales relative to conifers (A, B and C in Figure 1). The three groupings are supported by each of these genes in both observed and corrected spectra. The signal for each of these groupings, however, is generally increased as the result of Hadamard correction (compare the frequency values for splits A, B and C in panels I and II of Figure 3, 4 and 5). Split A, which supports the grouping of Gnetales and conifers, is supported by the strongest signal in all three genes. The other two splits (B and C) are also supported (generally with close frequencies). Neither of these signals is dramatically increased or decreased as the result of Hadamard correction (which corrects for multiple substitutions). Figure 6 shows the observed and Hadamard corrected spectra for the chloroplast gene *psaA*. The pattern of the signal/conflict did not change in the Hadamard correction. This pattern, however, is different from the pattern obtained from *rpa1, rpb1* and *rpc1*. A strong signal supporting grouping C (Gnetales and Pinaceae; Gnepines) is

present in both observed and Hadamard corrected spectra of *psa*A (Figure 6, I and II respectively). A weaker signal supporting grouping A (Gnetales and conifers together) is also present in the spectra of *psa*A. However, the signal supporting grouping B (Pinaceae and other conifers together) is not present in the observed spectrum of *psa*A and is only present with a very low frequency in the Hadamard corrected spectrum of this gene. Interestingly, the opposite pattern is seen in the spectrum of 18S (Figure 7): signal for grouping C (Gnepines) is very weak and signal for the other two groupings are present with high frequency in both observed and Hadamard corrected spectra. Figure 8 compares the observed and Hadamard corrected signal/conflict frequency values for three splits regarding the position of Gnetales and conifers (groupings A, B and C in Figure 1) for all 10 genes. Grouping of conifers and Gnetales is supported in the spectrum of a majority of the genes (both observed and Hadamard corrected). The highest frequency of support for this grouping is found in the spectrum of nuclear *rpa1, rpb1* and *rpc1* genes. Among the two opposing groupings, the B (Pinaceae and other coinfers together) and C (Gnepines), Gnepines grouping is abundant in the majority of genes and has a higher support in the spectrum of chloroplast and mitochondrial genes. The signal for grouping B is either not present or very weak in the spectrum of chloroplast and mitochondrial genes (Figure 8), but is present in nuclear genes.

Phylogenetic reconstructions

Phylogenies made using nuclear RNA polymerases data set

Figure 9 shows ML and MP phylogenetic trees made using combined data set of nuclear *rpa1*, *rpb1* and *rpc1* (hereafter called the RNAP data set). Based on the large

difference between NS values of 1+2 codon positions and third codon positions of these genes (see table 1), and the results of our evolutionary distance analyses (Table 2 and Figure 2 where synonymous sites and third codon positions of RNA polymerases showed large values, mostly above 1 substitution per site), we made phylogenetic trees with three data partitions: all codon positions, 1+2 codon positions and third codon positions. ML and MP trees made using all three codon positions (6723 sites, 3133 variable and 2104 parsimony-informative) both support gymnosperms as a monophyletic group (Figures 9A and B, respectively). The position of Gnetales is different in ML and MP trees. In ML they are nested within conifers and a sister group to Pinaceae (bootstrap value of 63). However, in MP Gnetales are the first branch of gymnosperms (with a bootstrap value of 100). ML and MP trees made using 1+2 codon positions (4482 sites, 1227 variable and 621 parsimony-informative) both support gymnosperms as a monophyletic group (Figures 9C and D, respectively). These trees also support Gnetales as a sister group to conifers (grouping A in Figure 1). However, in the ML tree Gnetales are grouped with Pinaceae (bootstrap value of 56; Figure 9C). In the MP tree, Pinaceae are grouped with other conifers (Figure 9D), but this grouping is not supported by bootstrapping (note that bootstrap values below 50 are not shown on the trees). ML and MP trees made using third codon positions (2241 sites, 1906 variable and 1483 parsimony-informative; Figures 9E and F) have different topologies. The ML tree supports the grouping of Gnetales with conifers (including Pinaceae; Figure 9E). However, the MP tree groups Gnetales as the first branch of gymnosperms (Figure 9F). This topology is identical to the MP tree of all three codon positions.

Phylogenetic trees from chloroplast data set

Figure 10 shows ML and MP phylogenetic trees made using combined data set of chloroplast genes *psaA*, *psaB*, *rbcL* and *atpB*. As mentioned above, based on the large difference between NS values of 1+2 codon positions and third codon positions of chloroplast genes (see Table 1) and also considering evolutionary distances (Figure 2A where a large number of distances for chloroplast genes are close to or above 1 substitution per site), we made phylogenetic trees with three data partitions: all codon positions, 1+2 codon positions and third codon positions. There is a clear disagreement between topologies from different data partitions.

The ML tree of all codon positions supports the Gnepines grouping (Figure 10A), but MP resolved Gnetales as the first branch of gymnosperms (Figure 10B). ML and MP trees of 1+2 codon positions both support the Gnepines grouping with high bootstrap support (bootstrap values of 97 and 96, respectively; Figures 10C and D, respectively). However, the ML tree from third codon positions groups Gnetales and conifers (not including Pinaceae; Figure 10E) and the MP tree groups Gnetales as the first branch of gymnosperms (Figure 10F).

Phylogenetic trees from mitochondrial data set

We made phylogenetic trees from all three codon positions of mitochondrial genes *atp*A and *cox1*. This data set was not partitioned into 1+2 and third codon positions because the NS analysis did not show a large difference in NS values of different partitions of mitochondrial genes (Table 1). Furthermore, evolutionary distances for mitochondrial genes were mostly under 1 substitution per site (Figure 2A). Interestingly,

ML and MP produced identical trees (Figures 11A and B, respectively). The grouping of Gnetales and conifers was supported in this topology, but *Pinus* was grouped outside this conifers-Gnetales clade, with cycads. In addition, gymnosperms are not monophyletic in these trees (Figures 11A and B).

Phylogenetic trees from 18S rRNA

ML and MP trees of 18S rRNA (Figures 11C and D, respectively) grouped Gnetales with conifers (including Pinaceae) with high bootstrap support (bootstrap values of 100 and 97, respectively). However, the position of *Pinus* is different in ML and MP trees and angiosperms are not resolved in MP (Figures 11C and D, respectively).

Phylogenetic analysis using total evidence

A data set containing all 10 genes (16223 sites, 6327 variable and 3980 parsimony-informative) was used in ML and MP analyses (Figures 12A and B, respectively). The ML tree supports the grouping of Gnetales and conifers with a bootstrap value of 100. In this tree Gnetales are further grouped with *Pinus* with a bootstrap support of 89. In the MP tree, however, Gnetales are the first branch of gymnosperms. The position of Ginkgo is also different in ML and MP trees. In the ML tree, Ginkgo is the sister group of conifers and Gnetales but in MP Ginkgo is the sister group of cycads.

To study the effect of removing third codon positions of the RNAP and chloroplast genes (see above), we made phylogenetic trees from combined data set of 1+2 codon positions of RNAP and chloroplast genes and all codon positions of mitochondrial

genes and 18S rRNA genes. Interestingly, This combined data set produced a single topology in ML and MP (Figures 13A and B, respectively). In this topology, Gnetales are grouped with conifers, as a sister group to Pinaceae, with bootstrap value of 100 both in ML and MP.

We also made a MP phylogenetic tree using translated amino acid sequences of all protein-coding genes. This tree (Figure 14) supports the grouping of Gnetales and conifers and the grouping of Gnetales and Pinaceae. However, Ginkgo is positioned with cycads in this tree.

Discussion

Beside the complexity of addressing phylogenetic relationship of seed plants, the choice of data sets and gene partitions are the major causes of confusion and contradictory results in recent phylogenetic reconstructions of seed plants (e.g., Magallón and Sanderson 2002; Soltis, Soltis, and Zanis 2002). To address this, we used normalized splits and evolutionary distances to evaluate data sets. The NS method allows comparing the noise in genes and gene partitions without building a phylogenetic tree. Measuring distances allows identifying sites that might be a concern for saturation. Our results from NS analysis and evolutionary distances are in line with what others have suggested with conventional phylogenetic methods (Magallón and Sanderson 2002; Soltis, Soltis, and Zanis 2002): mainly that the third codon positions are noisy at the divergence level of seed plant phylogeny. However, these studies only examined different partitions of chloroplast genes.

Our results with spectral analysis strongly support the proposed (by molecular studies) view on the closeness of Gnetales and conifers (Donoghue and Doyle 2000). The signal supporting this grouping is present in a vast majority of genes (Figure 8). Gnepines is also supported by all protein-coding genes. The frequency of support, however, is different among different genes. This is in fact in line with the argument that different genes have different phylogenetic resolution for a given divergence level. The strong supports for Gnepines from the spectrum of mitochondrial genes *atpA* and *cox1* , and chloroplast genes *psaA* and *psbB* is in line with phylogenetic studies by Bowe, Coat, and dePamphilis (2000) and Magallón and Sanderson (2002), respectively. However, there is also support for the grouping of Pinaceae with other conifers in a number of genes (namely nuclear RNA polymerases and 18S rRNA). We believe this could be in fact a historical signal and not contradictory to the Gnepines grouping (although it might mislead a phylogeny depending on gene, partition or method of analysis). The ancestral signal, which leads to this grouping, might not be present in Gnetales, possibly due to their faster rate of evolution, but is present in Pinaceae and other conifers. This fast rate of evolution (relative to other seed plants) in Gnetales is a well known issue, and documented in our phylogenetic trees (Figures 9-14), as well as trees in several other phylogenetic studies (e.g., Bowe, Coat, and dePamphilis 2000; Chaw et al. 2000) by a clear long branch leading to Gnetales. Our results further suggests that a single-gene phylogeny can be misleading for reconstructing evolutionary history of seed plants. For example the spectrum of *psaA* and 18S rRNA clearly suggest that using each gene will lead to two different positions for the Gnetales (Figures 6 and 7, respectively).

Conventional phylogenetic reconstructions using ML and MP further supports the result we obtained with spectral analysis. In this study we introduced a new data set from nuclear protein coding genes *rpa1*, *rpb1* and *rpc1*. This is the first extensive multi-gene data set of nuclear protein-coding genes to address seed plant phylogeny. This data set supports the Gnepines hypothesis in ML analysis of all codon positions and 1+2 codon positions. We summarized the results obtained in phylogenetic reconstructions regarding the position of Gnetales to conifers (Table 3). All ML trees support the grouping of Gnetales and conifers (grouping A; Table 3) with generally a high bootstrap support. Gnepines is also supported by a majority of data sets (grouping C; Table 3). Grouping of *Pinus* and other conifers (grouping B; Table 3) is only supported by ML using 18S rRNA and third codon positions of RNAP. The variation in the position of Gnetales is higher in MP analysis of different data sets (Table 3). In MP the only data set that resolves the same topology as the abundant topology in ML (A and C groupings in Figure 1), is the data set of 1+2 codon positions of chloroplast genes. Interestingly, this data set evolves very slowly (chloroplast genes have the lowest number of non-synonymous substitutions; and the 1+2 codon positions of chloroplast genes are the slowest evolving data partition, see Table 2 and Figure 2B). Our results confirm the results of studies that originally proposed the Gnepines hypothesis (Bowe, Coat, and dePamphilis 2000; Chaw et al. 2000) and studies that further found support for it (i.e. Magallón and Sanderson 2002; Soltis, Soltis, and Zanis 2002). Other studies that did not support Gnepines either used single genes or small data sets (Schmidt and Schneider-Poetsch 2002), or included third codon positions (Rydin, Källersjö, and Friis 2002; Schmidt and Schneider-Poetsch 2002). The problem with using one gene to address a complex phylogenetic issue, such as the

seed plant phylogeny, is often the lack of phylogenetic information. Authors, however, have tried to incorporate third codon positions to make use of the information contained in these sites. This assumption is simply false for a complex phylogeny such as the phylogeny of seed plants, with a depth of several hundred million years, because third codon positions have incorporated multiple substitutions in each lineage and either have lost their phylogenetic signal or the signal is highly influenced by noise. Several studies have mentioned the problem with third codon positions (Meyer 1994; Swofford et al. 1996; Yang 1998; Xia et al. 2003). This problem is especially important when a method sensitive to multiple substitutions (i.e., MP) is used for reconstructing phylogenies (the problem of long branch attraction; Felsenstein 1978). Combining datasets have the advantage of increasing the signal. And several recent studies on plant phylogeny have shown this advantage (e.g. Qiu et al. 1999; Soltis, Soltis, and Chase 1999). However, when the noisy data partitions are not excluded, the signal is still influenced by noise and can be distorted or masked by it (Chapter 3). One way of reducing the noise is to use a method of analysis with realistic models of nucleotide evolution and estimating the parameters of the model from the data set. Maximum likelihood methods provide this possibility. We clearly showed that using a ML approach (with model and parameters estimated from data set) different data sets converge on a uniform topology for seed plants. Another way for reducing the noise is to use data (data partitions) that is relatively free of noise. Using amino acid sequences is an example of this strategy (although it may require a larger number of amino acids to resolve a topology). Interestingly, the MP tree of amino acid sequences of all protein-coding genes grouped Gnetales and Pinaceae (Figure 14). Another example is to use a method like NS (without considering a model of

nucleotide evolution) or evolutionary distances (based on a model of nucleotide substitutions) to evaluate genes/data partitions. This strategy has the advantage of increasing the signal for all branches of the tree (by adding the signal in different genes) and decreasing the noise (by removing data partitions that are noisy). This should lead to a single topology in ML and MP. We therefore combined 1+2 codon positions of chloroplast and nuclear genes and all codon positions of mitochondrial genes and 18S rRNA in a data set and built phylogenies with this data set in ML and MP. A single topology was obtained with high bootstrap support in both ML and MP (Figures 13A and B). In this topology, which was identical to the topology obtained by ML using total evidence (all genes, all codon positions), Gnetales are nested within conifers as a sister group to Pinaceae.

References

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

Chase, M., D. Soltis, R. Olmstead, D. Morgan, and D. Les, et al. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. Ann. Mo. Bot. Gard. 80:526-580.

Chaw, S. M., C. L. Parkinson, Y. Cheng, T. M. Vincent, and J. D. Palmer. 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant

gymnosperms and origin of Gnetales from conifers. Proc. Natl. Acad. Sci. USA 97:4086-4091.

Crane, P. 1985. Phylogenetic analysis of seed plants and the origin of angiosperms. Ann. Mo. Bot. Gard. 72:716-793.

Donoghue, M. J., and J. A. Doyle. 2000. Seed plant phylogeny: Demise of the anthophyte hypothesis? Curr. Biol. 10:R106-109.

Doyle, J., and M. Donoghue. 1992. Fossils and seed plant phylogeny reanalyzed. Brittonia 44:89-106.

Doyle, J., and M. Donoghue. 1986. Seed plant phylogeny and the origin of angiosperms: an experimental cladistic approach. Bot. Rev. 52:321-431.

Doyle, J., M. Donoghue, and E. Zimmer. 1994. Integration of morphological and ribosomal RNA data on the origin of angiosperms. Ann. Mo. Bot. Gard. 81.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401-410.

Goremykin, V., V. Bobrova, J. Pahnke, A. Troitsky, A. Antonov, and W. Martin. 1996. Noncoding sequences from the slowly evolving chloroplast inverted repeat in addition to rbcL data do not support gnetalean affinities of angiosperms. Mol. Biol. Evol. 13:383-396.

Hamby, R., and E. Zimmer. 1992. Ribosomal RNA as a phylogenetic tool in plant systematics. Pp. 50-91 *in* P. Soltis, D. Soltis, and Doyle JJ, eds. Molecular Systematics of Plants. Chapman & Hall, New York.

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

Hendy, M. D., and D. Penny. 1993. Spectral analysis of phylogenetic data. J. classification **10**:5-24.

Hendy, M. D., D. Penny, and M. A. Steel. 1994. A discrete Fourier analysis for evolutionary trees. Proc. Natl. Acad. Sci. USA **91**:3339-3343.

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

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

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

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111-120.

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17**:1244-1245.

Lento, G. M., R. E. Hickson, G. K. Chambers, and D. Penny. 1995. Use of spectral analysis to test hypotheses on the origin of pinnipeds. Mol. Biol. Evol. **12**:28-52.

Lockhart, P. J., D. Penny, and A. Meyer. 1995. Testing the phylogeny of swordtail fishes using split decomposition and spectral analysis. J. Mol. Evol. 41:666-674.

Magallón, S., and M. J. Sanderson. 2002. Relationships among seed plants inferred from highly conserved genes: sorting conflicting phylogenetic signals among ancient lineages. Am. J. Bot. **89**:1991-2006.

Meyer, A. 1994. Shortcomings of the cytochrome b gene as a molecular maker. Trends Ecol. Evol. **9**:278-280.

Nickerson, J., and G. Drouin. 2003. The sequence of the largest subunit of RNA polymerase II is a useful maker for inferring seed plant phylogeny. Mol. Phylogenet. Evol., in press

Pamilo, P., and N. Bianchi. 1993. Evolution of the Zfx and Zfy genes : rates and interdependence between the genes. Mol. Biol. Evol. **10**:271-281.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14**:817-818.

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

Rydin, C., M. Källersjö, and E. M. Friis. 2002. Seed plant relationships and the systematic position of Gnetales based on nuclear and chloroplast DNA: Conflicting data, rooting problems, and the monophyly of conifers. Int. J. Plant Sci. **163**:197-214.

Sanderson, M. J., M. F. Wojciechowski, J. M. Hu, T. S. Khan, and S. G. Brady. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol. Biol. Evol. **17**:782-797.

Schmidt, M., and H. A. W. Schneider-Poetsch. 2002. The evolution of gymnosperms redrawn by phytochrome genes: The Gnetatae appear at the base of the gymnosperms. J. Mol. Evol. **54**:715-724.

Soltis, D. E., P. S. Soltis, and M. Zanis. 2002. Phylogeny of Seed Plants Based on Evidence from Eight Genes. Am. J. Bot. **89**:1670-1681.

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

Stefanovic, S., M. Jager, J. Deutsch, J. Broutin, and M. Masselot. 1998. Phylogenetic relationships of conifers inferred from partial 28S rRNA gene sequences. Am. J. Bot. **85**:688-697.

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

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

Swofford, D. L. 2000. PAUP*.Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, MA.

Swofford, D. L., G. L. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407-514 *in* D. M. Hillis, C. Morowitz, and M. B.K., eds. Molecular Systematics. Sinauer, Sunderland, MA.

Winter, K. U., A. Becker, T. Munster, J. T. Kim, H. Saedler, and G. Theissen. 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.

Xia, X., Z. Xie, M. Salemi, L. Chen, and Y. Wang. 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol. **26**:1-7.

Yang, Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. **47**:125-133.

Table 1. NS values for different genes and gene partitions

Data set	Sites	Splits D	NS-D	Splits H	NS-H	Variable	Parsimony
rpa1	1596	180	0.113	282	0.177	920	577
rpa1 (1+2)	1064	90	0.085	149	0.140	439	224
rpa1 (third)	532	147	0.276	N/A	N/A	481	353
rpc1	1638	241	0.147	401	0.245	867	564
rpc1 (1+2)	1092	101	0.092	176	0.161	369	185
rpc1 (third)	546	205	0.375	N/A	N/A	498	379
rpb1	3039	417	0.137	716	0.236	1346	963
rpb1 (1+2)	2026	121	0.060	168	0.083	419	212
rpb1 (third)	1013	376	0.371	N/A	N/A	927	751
rbcL	1169	140	0.120	211	0.180	367	227
rbcL (1+2)	780	39	0.050	46	0.059	80	38
rbcL (third)	389	124	0.319	706	1.815	287	189
psaA	2112	205	0.097	328	0.155	719	456
psaA (1+2)	1407	66	0.047	82	0.058	192	107
psaA (third)	705	182	0.258	562	0.797	527	349
psbB	1347	141	0.105	258	0.192	480	305
psbB (1+2)	898	46	0.051	57	0.063	141	72
psbB (third)	449	129	0.287	554	1.234	339	233
atpB	1146	109	0.095	218	0.190	380	243
atpB (1+2)	764	35	0.046	49	0.064	89	47
atpB (third)	382	95	0.249	818	2.141	291	196
atpA	1104	88	0.080	120	0.109	398	215
atpA (1+2)	736	50	0.068	66	0.090	157	85
atpA (third)	368	65	0.177	156	0.424	241	130
cox1	1392	112	0.080	157	0.113	551	263
cox1 (1+2)	928	70	0.075	92	0.099	256	128
cox1 (third)	464	64	0.138	125	0.269	295	135
18S rRNA	1670	105	0.063	171	0.102	293	165

Table 2. Overall mean of number of synonymous and non-synonymous substitutions for each protein coding gene and data set. The mean values were calculated using all the lineages used in this study.

Gene	Ks±SE	Ka±SE
rpa1	1.063±0.045	0.189±0.012
rpb1	1.223±0.036	0.060±0.004
rpc1	1.096±0.050	0.122±0.008
PsaA	0.504±0.020	0.032±0.003
PsabB	0.525±0.024	0.040±0.004
rbcL	0.462±0.026	0.020±0.003
atpB	0.530±0.034	0.031±0.005
cox1	0.341±0.018	0.084±0.006
atpA	0.363±0.020	0.073±0.007
Data set		
RNAP	1.188±0.025	0.097±0.003
Chloroplast	0.502±0.012	0.031±0.002
Mitochondrial	0.341±0.014	0.081±0.005

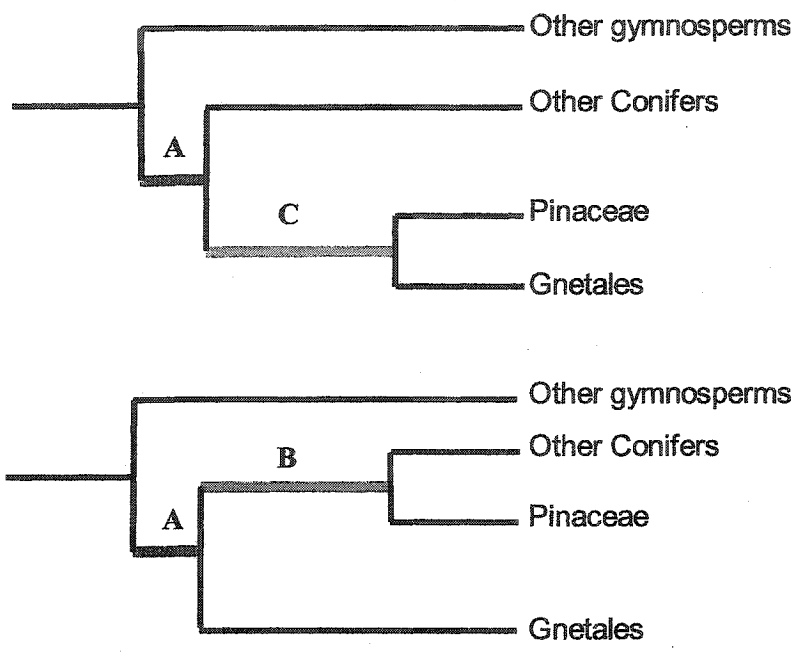
Note: Ks=synonymous, Ka=non-synonymous, SE=standard error measured by bootstrapping (100 replicates) Values calculated using Pamilo-Bianchi-Li method (Pamilo and Bianchi 1993)

Table 3. Support for different groupings regarding the relative position of Gnetales and conifers in different phylogenetic trees.

Data set	ML	MP
RNAP	A(98), C(63)	A(100), B(95)
RNAP (1+2)	A(98), C(56)	A(76), B(<50)
RNAP (third)	A(89), B(65)	D(100), B(91)
Chloroplast	A(94), C(70)	D(97), B(96)
Chloroplast (1+2)	A(94), C(97)	A(85), C(96)
Chloroplast (third)	A(71)	B(96)
Mitochondrial	A(<50)	A(61)
18S rRNA	A(69), B(57)	A(97), B(<50)
All genes	A(100), C(89)	D(93), B(77)
All genes (excluding third bases of RNAP and chloroplast genes)	A(100), C(100)	A(100), C(100)
Amino acid translation of protein coding genes	N/A	A(86), C(100)

Note: ML= maximum likelihood; MP= maximum parsimony; numbers in parentheses are bootstrap supports; A: (Gnetales+conifers) (other gymnosperms); B: (conifers including Pinaceae) (Gnetales +other gymnosperms); C: (Gnetales + Pinaceae) (other conifers +other gymnosperms); D = Gnetales are the first branch of gymnosperms.

Figure 1. Schematic trees, and their split representation, showing three different groupings related to the relationship of Gnetales and conifers



A: (Gnetales+conifers) (other gymnosperms)

B: (conifers including Pinaceae) (Gnetales +other gymnosperms)

C: (Gnetales + Pinaceae) (other conifers +other gymnosperms)

Figure 2. Pairwise distances of third (A) and 1+2 codon positions (B) of RNA polymerases, chloroplast and mitochondrial data sets. Distances were estimated using K2P + Γ model (α = 0.5). Pairwise distance values for each gene are represented based on alphabetical order of species names (for example the distance of Amborella-cycas is plotted before the distance of Amborella-Ephedra and so on). Distances involving Gnetales species are shown in black.

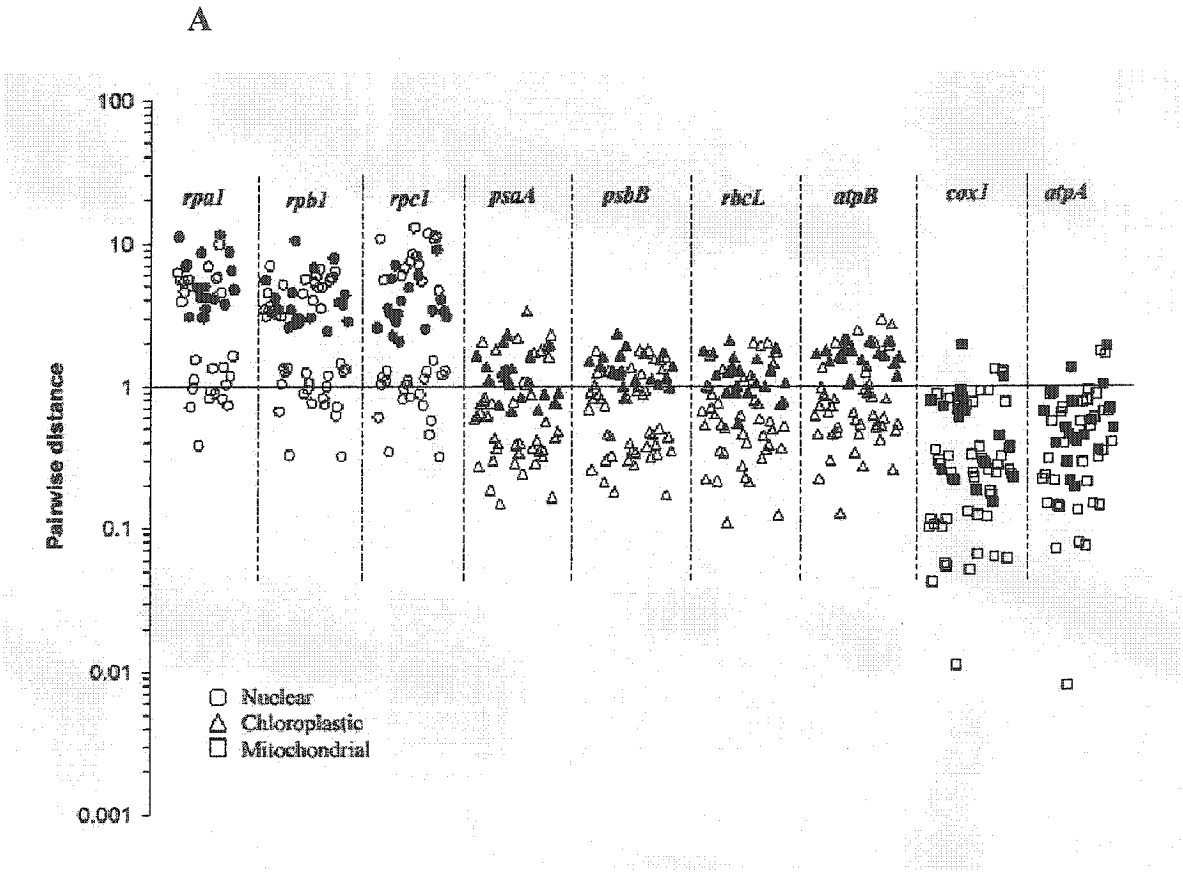


Figure 2. *Continued*

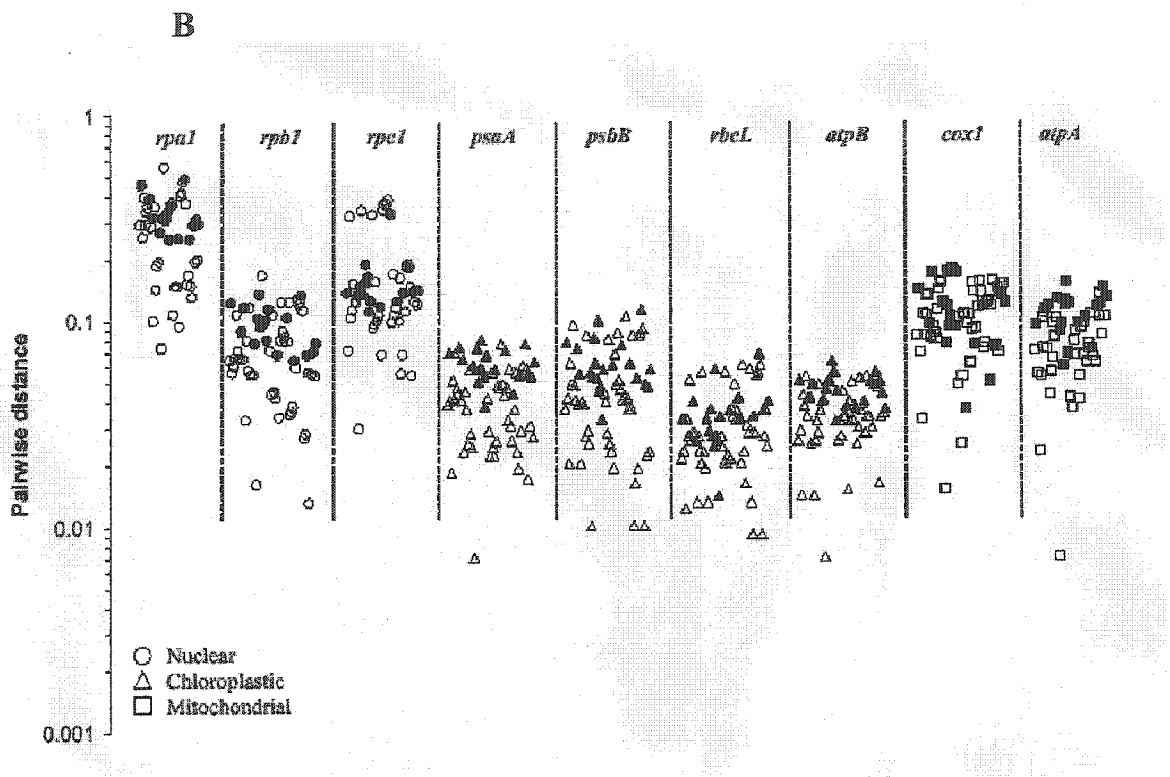
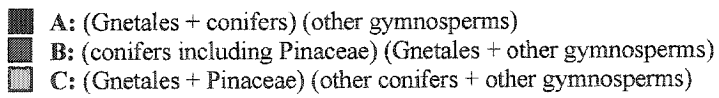
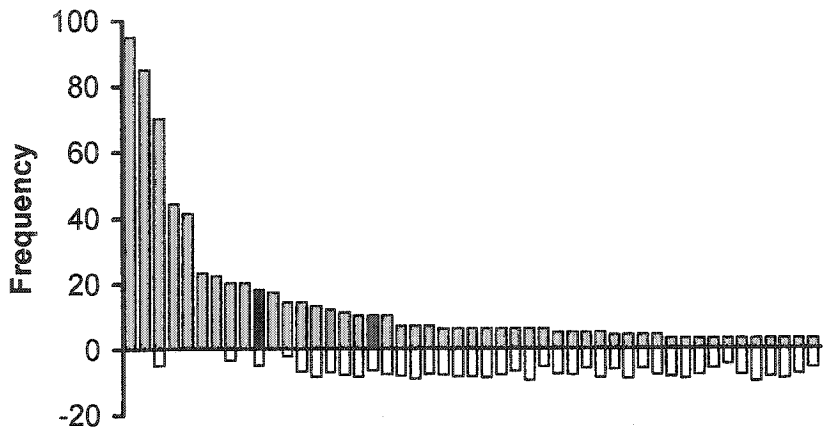


Figure 3. Observed and Hadamard-corrected spectrum of *rpa1*. This analysis is done using the alignment of 9 gymnosperm taxa. Support and conflict values for each split are shown as positive and negative bars, respectively. Support values for three splits regarding the relative position of Gnetales to conifers are color-coded. Note that 8 singleton splits (a single taxa separated from others) do not have conflict values.



I) Observed



II) Hadamard-corrected

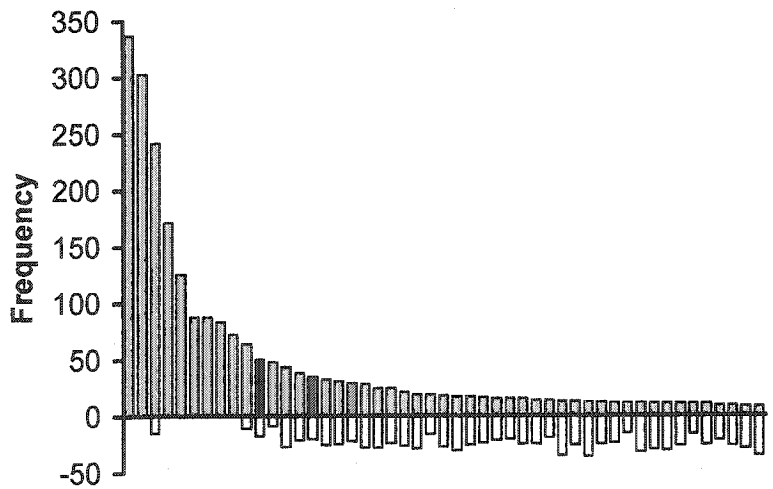
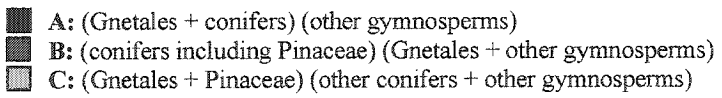
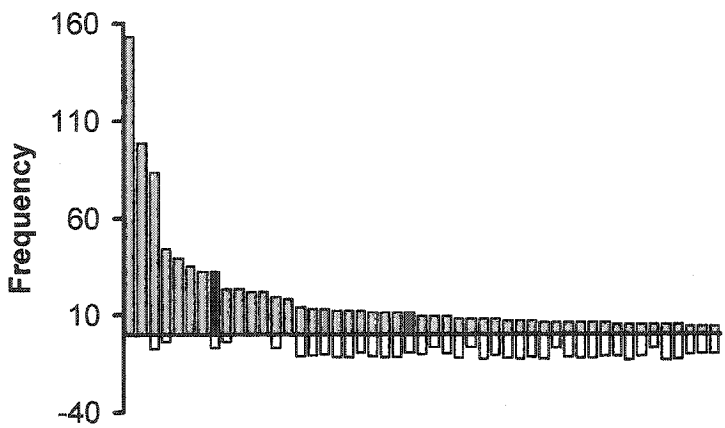


Figure 4. Observed and Hadamard-corrected spectrum of *rpb1*. This analysis is done using the alignment of 9 gymnosperm taxa. Support and conflict values for each split are shown as positive and negative bars, respectively. Support values for three splits regarding the relative position of Gnetales to conifers are color-coded. Note that 8 singleton splits (a single taxa separated from others) do not have conflict values.



I) Observed



II) Hadamard-corrected

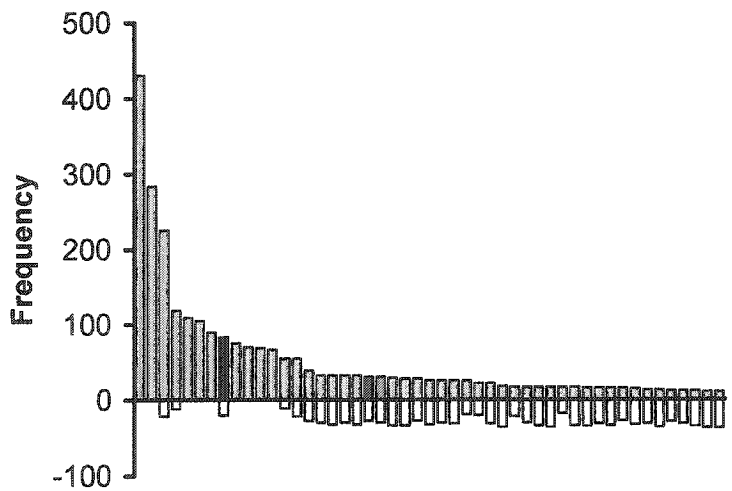
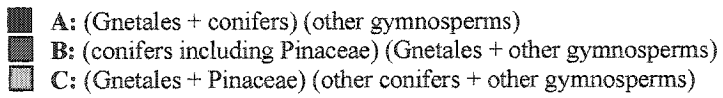
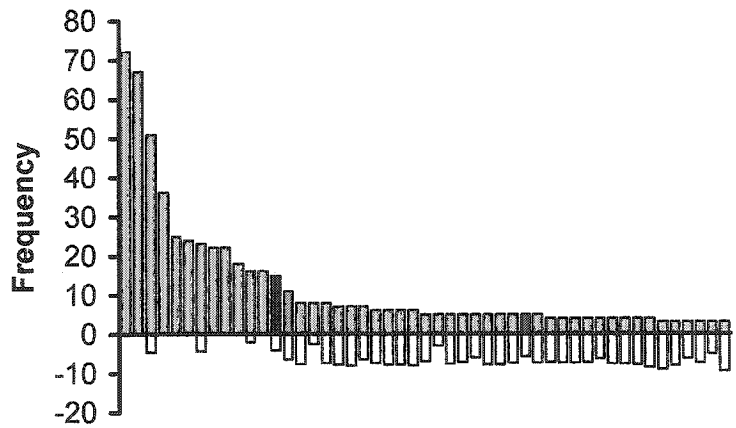


Figure 5. Observed and Hadamard-corrected spectrum of *rpc1*. This analysis is done using the alignment of 9 gymnosperm taxa. Support and conflict values for each split are shown as positive and negative bars, respectively. Support values for three splits regarding the relative position of Gnetales to conifers are color-coded. Note that 8 singleton splits (a single taxa separated from others) do not have conflict values.



I) Observed



II) Hadamard-corrected

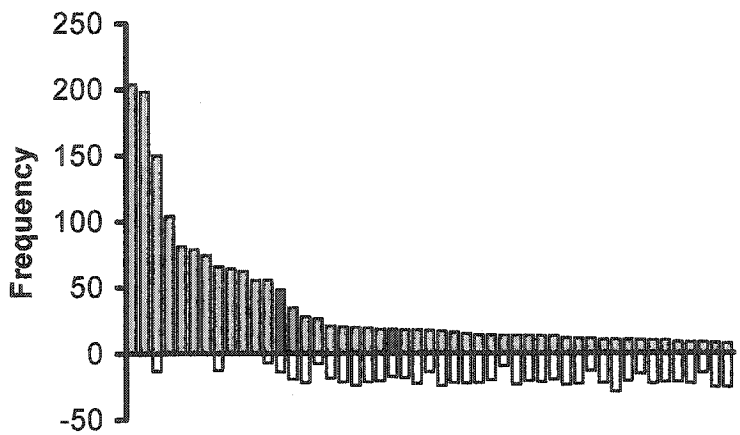


Figure 6. Observed and Hadamard-corrected spectrum of *psaA*. This analysis is done using the alignment of 9 gymnosperm taxa. Support and conflict values for each split are shown as positive and negative bars, respectively. Support values for three splits regarding the relative position of Gnetales to conifers are color-coded. Note that 8 singleton splits (a single taxa separated from others) do not have conflict values.

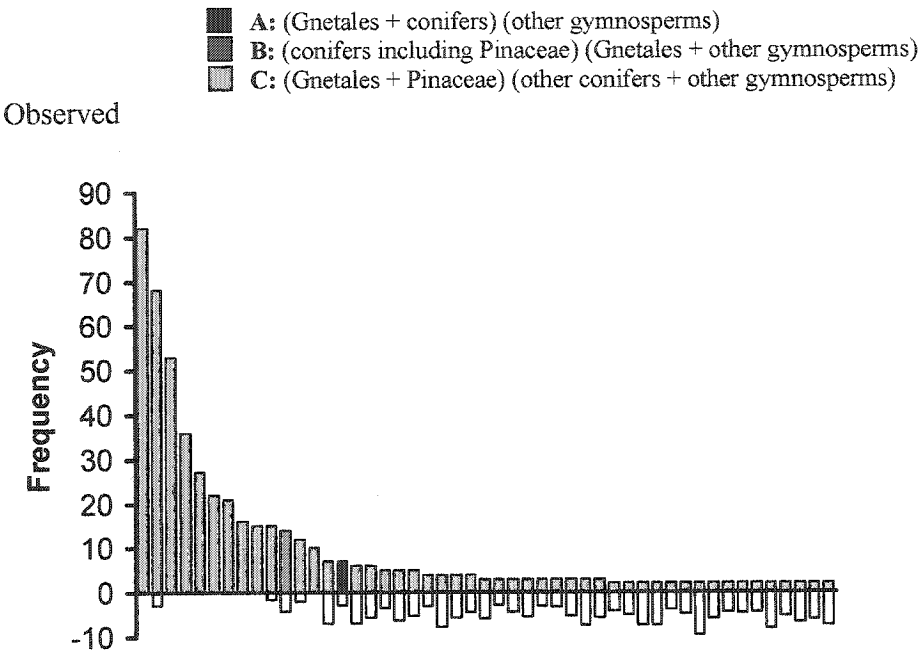
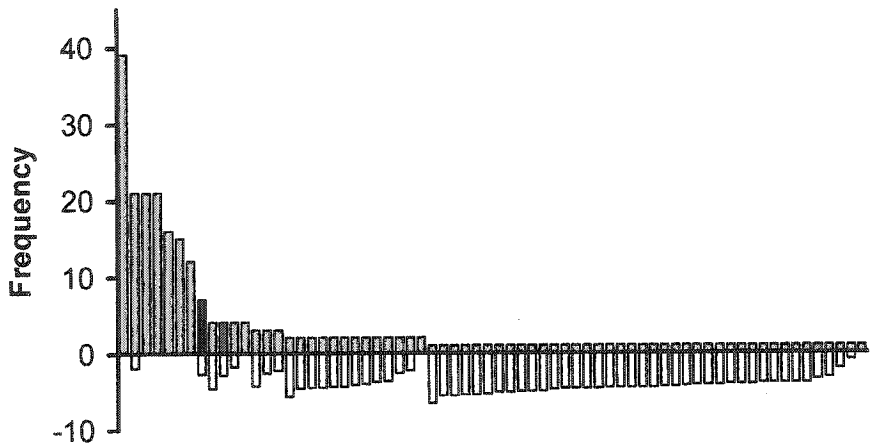


Figure 7. Observed and Hadamard-corrected spectrum of 18S rRNA. This analysis is done using the alignment of 9 gymnosperm taxa. Support and conflict values for each split are shown as positive and negative bars, respectively. Support values for three splits regarding the relative position of Gnetales to conifers are color-coded. Note that 8 singleton splits (a single taxa separated from others) do not have conflict values.

A: (Gnetales + conifers) (other gymnosperms)
B: (conifers including Pinaceae) (Gnetales + other gymnosperms)
C: (Gnetales + Pinaceae) (other conifers + other gymnosperms)

I) Observed



II) Hadamard-corrected

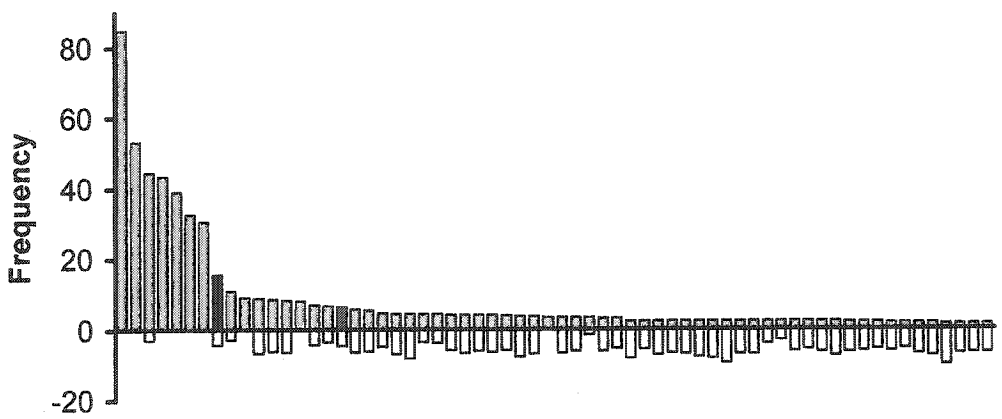
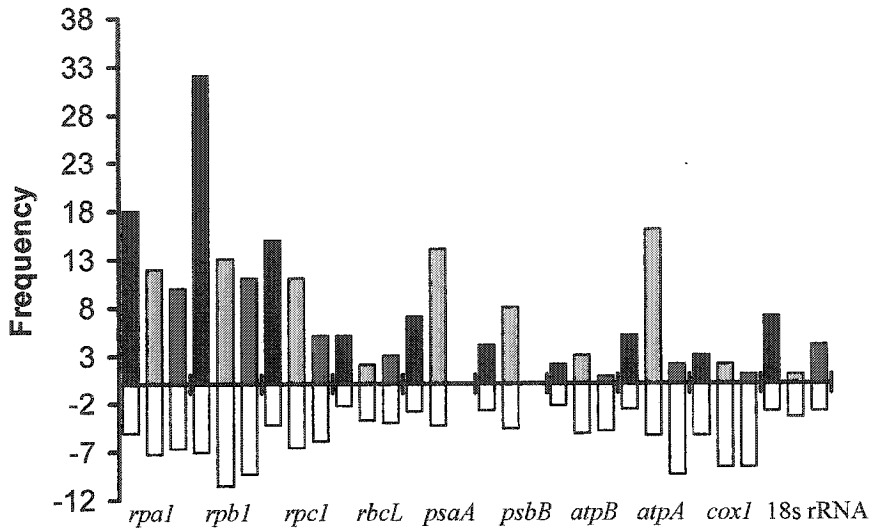


Figure 8. Support/conflict from spectrum of all genes for relative position of Gnetales to conifers. This analysis is done using the alignment of 9 gymnosperm taxa. Support and conflict values for each split are shown as positive and negative bars, respectively.

A: (Gnetales + conifers) (other gymnosperms)
B: (conifers including Pinaceae) (Gnetales + other gymnosperms)
C: (Gnetales + Pinaceae) (other conifers + other gymnosperms)

I) Observed



II) Hadamard-corrected

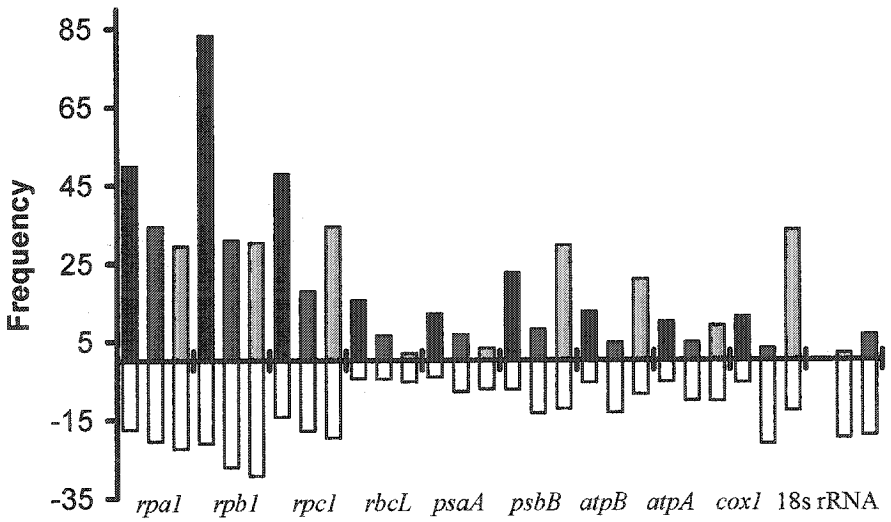
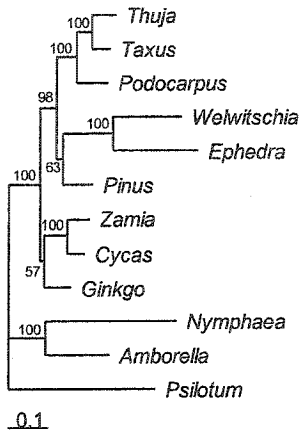
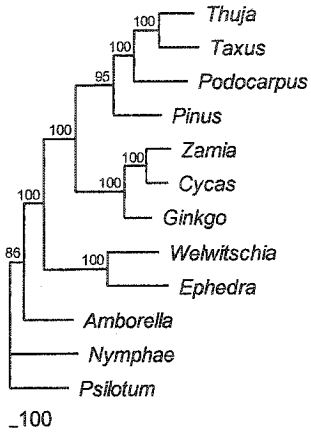


Figure 9. Phylogenetic trees from RNA polymerases data set. Trees are made using Maximum Likelihood (ML) and Maximum Parsimony. Bootstrap supports (100 replicates for ML and 1000 replicates for MP) are shown on each branch. Note that bootstrap values less than 50 are not shown.

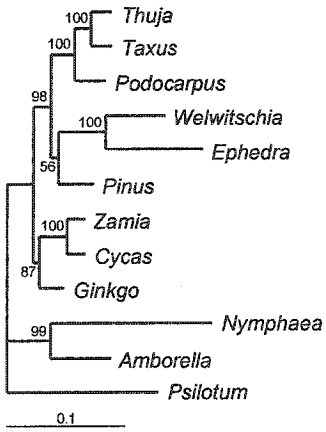
A) ML all codon positions



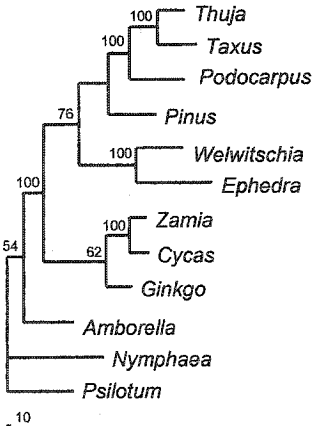
B) MP all codon positions



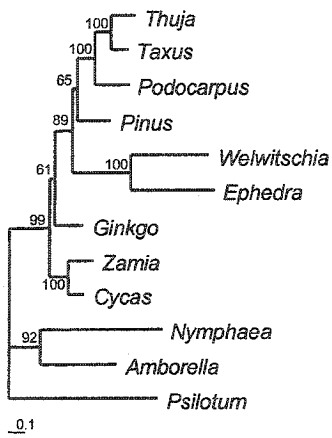
C) ML 1+2 codon positions



D) MP 1+2 codon positions



E) ML third codon position



F) MP third codon position

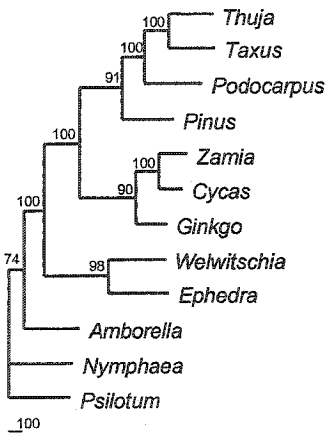
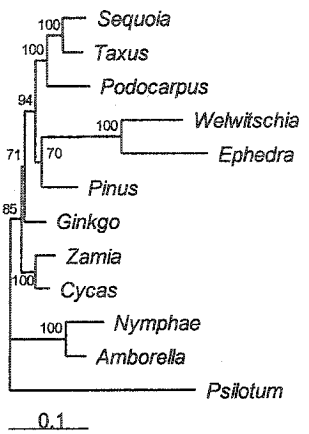
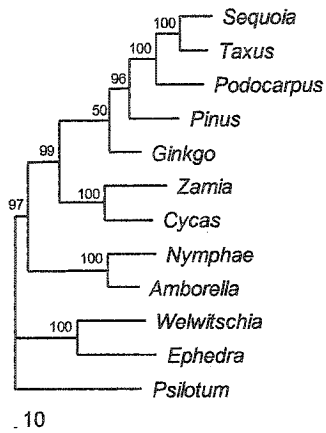


Figure 10. Phylogenetic trees from chloroplast genes data set. Trees are made using Maximum Likelihood (ML) and Maximum Parsimony. Bootstrap supports (100 replicates for ML and 1000 replicates for MP) are shown on each branch. Note that bootstrap values less than 50 are not shown.

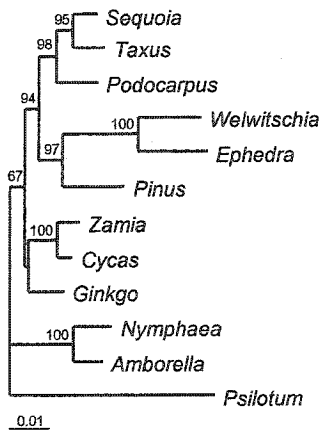
A) ML all codon positions



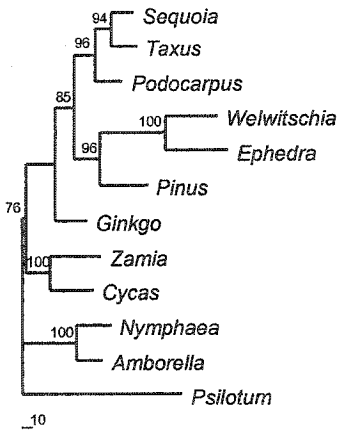
B) MP all codon positions



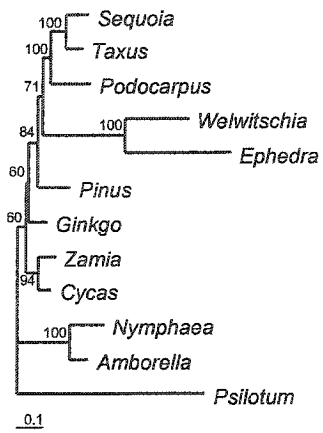
C) ML 1+2 codon positions



D) MP 1+2 codon positions



E) ML third codon position



F) MP third codon position

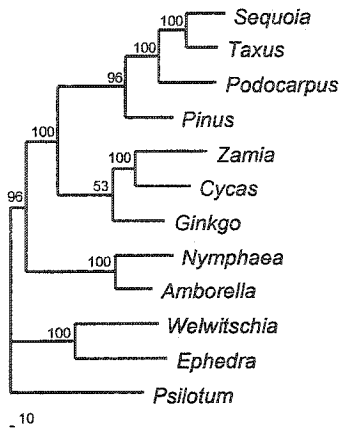


Figure 11. Phylogenetic trees from mitochondrial data set (A, B) and 18S rRNA (C, D). Trees are made using Maximum Likelihood (ML) and Maximum Parsimony. Bootstrap supports (100 replicates for ML and 1000 replicates for MP) are shown on each branch. Note that bootstrap values less than 50 are not shown.

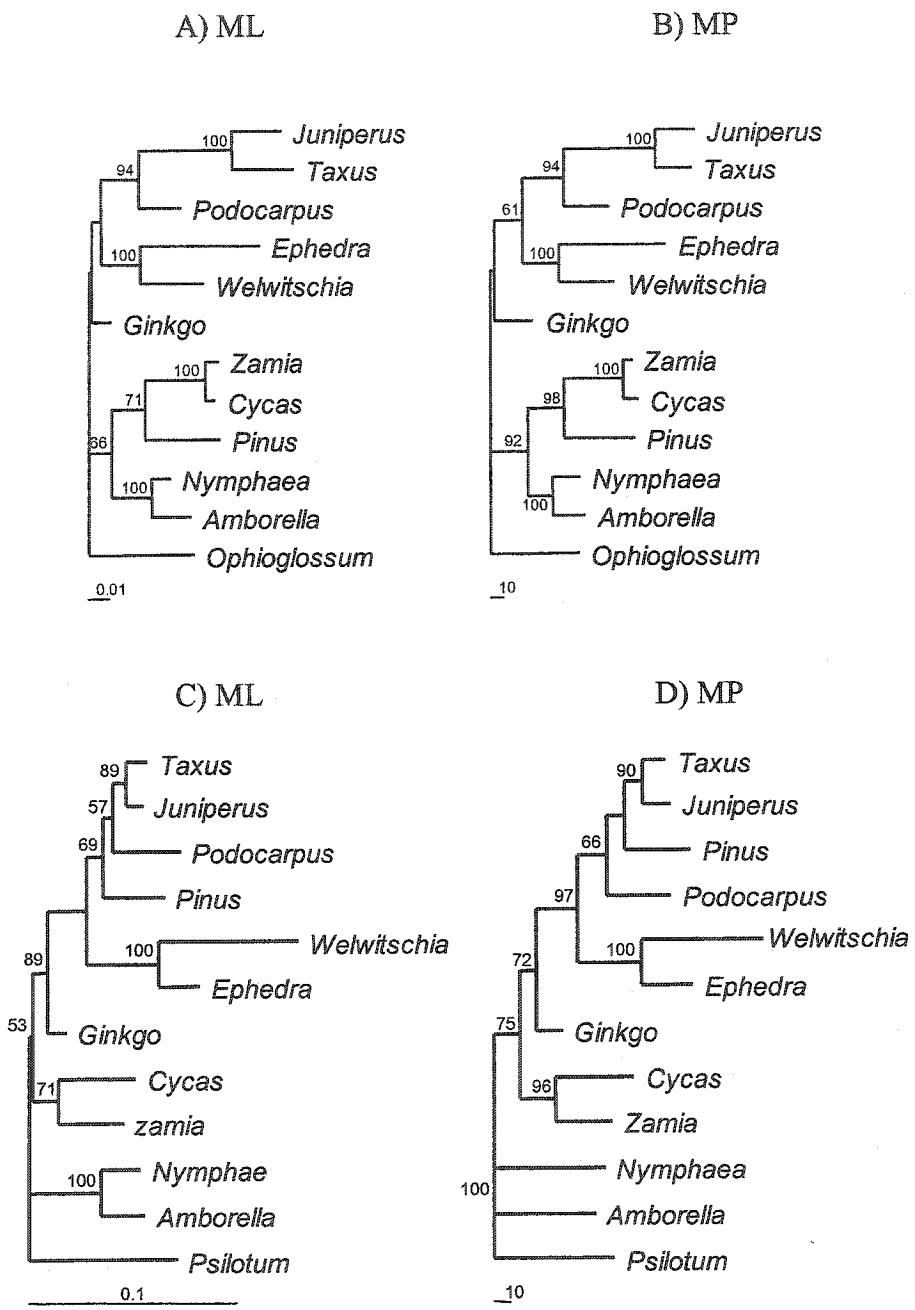


Figure 12. Phylogenetic trees from combined data set of all genes. Trees are made using Maximum Likelihood (ML) and Maximum Parsimony. Bootstrap supports (100 replicates for ML and 1000 replicates for MP) are shown on each branch.

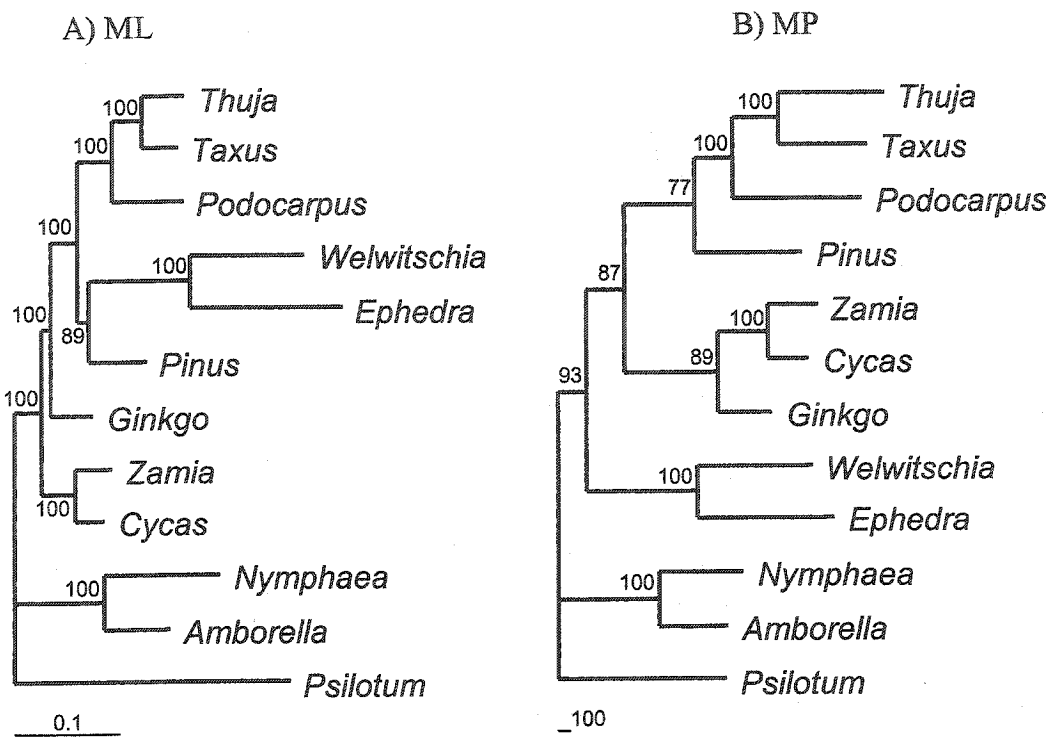


Figure 13. Phylogenetic trees from combined data set of 1+2 codon positions of RNAP and chloroplast genes and all codon positions of mitochondrial genes and 18S rRNA. Trees are made using Maximum Likelihood (ML) and Maximum Parsimony. Bootstrap supports (100 replicates for ML and 1000 replicates for MP) are shown on each branch.

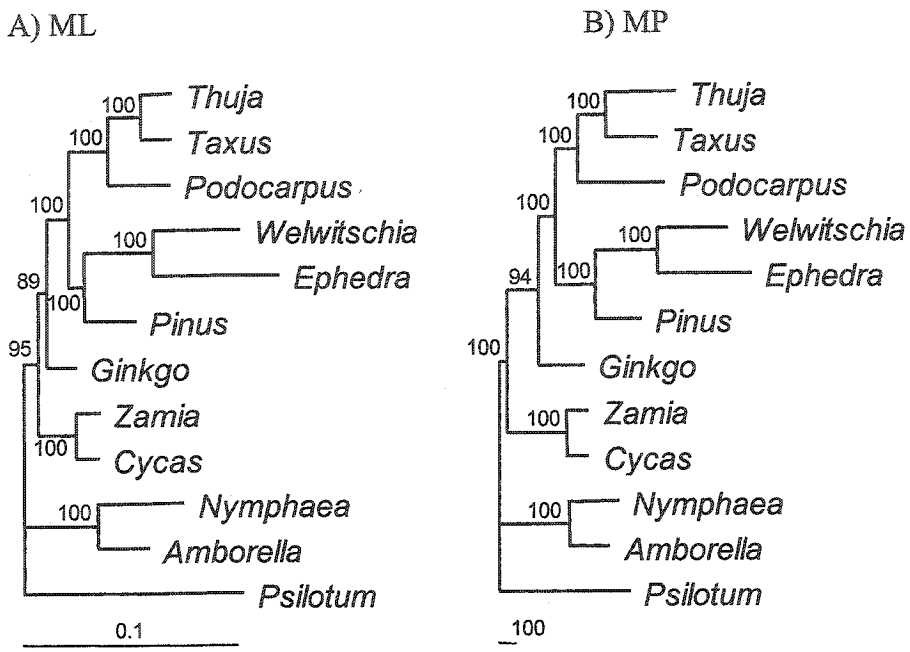
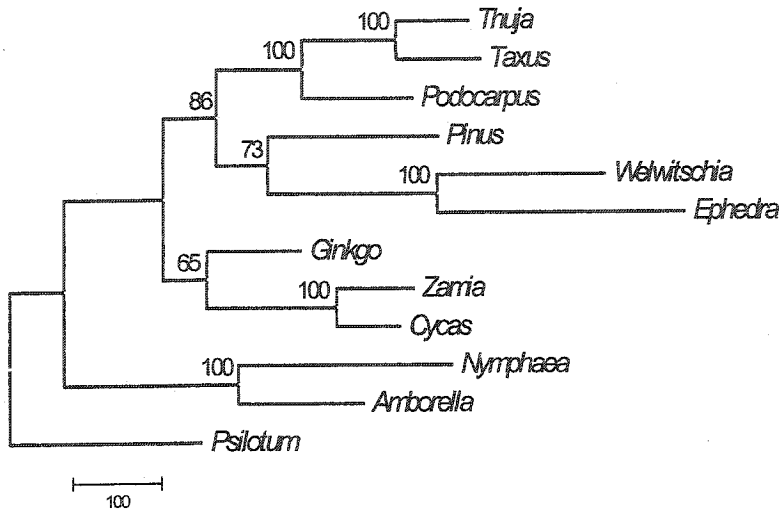


Figure 14. Maximum parsimony phylogenetic tree from translated amino acid sequences of all protein-coding genes. Bootstrap supports (1000 replicates) are shown on each branch.



Appendix 1. List of plant species and GenBank accession numbers for genes used in this study.

Family	Species	G.B.#		
		rpa1	*rpb1*	*rpc1*
Amborellaceae	*Amborella trichopoda*	AY490544	AF519541	N/A
Cupressaceae	*Thuja occidentalis*	N/A	AY490550	AY490557
Cycadaceae	*Cycas revoluta*	AY490547	AF519535	AY490561
Ephedraceae	*Ephedra viridis*	AY490543	AY490554	AY490556
Ginkgoaceae	*Ginkgo biloba*	AY490549	AY490553	AY490563
Nymphaeaceae	*Nymphaea odorata*	AY490545	AF519540	AY490565
Pinaceae	*Pinus nigra*	AY490546	AF519536	AY490560
Podocarpaceae	*Podocarpus macrophyllus*	AY490540	AY490552	AY490540
Psilotaceae[1]	*Psilotum nudum*	N/A	AF519542	AY490564
Taxaceae	*Taxus canadensis*	AY490541	AY490551	AY490558
Welwitschiaceae	*Welwitschia mirabilis*	AY490542	AF519537	AY490555
Zamiaceae	*Zamia muricata*	AY490548	AF519534	AY490562

	rbcL		*AtpB*	
Family	**Species**	**G.B.#**	**Species**	**G.B.#**
Amborellaceae	*Amborella trichopoda*	L12628	*Amborella trichopoda*	AF235041
Cupressaceae	*Thuja plicata*	AF127428	*Thuja plicata*	AY056545
Cycadaceae	*Cycas revoluta*	AY056556	*Cycas revoluta*	AY056518
Ephedraceae	*Ephedra californica*	AY056569	*Ephedra andina*	AY056538
Ginkgoaceae	*Ginkgo biloba*	D10733	*Ginkgo biloba*	AJ235481
Nymphaeaceae	*Nymphaea odorata*	M77034	*Nymphaea odorata*	AJ235544
Pinaceae	*Pinus thunbergii*	D17510	*Pinus thunbergii*	D17510
Podocarpaceae	*Podocarpus totara*	AF307931	*Podocarpus milanjianus*	AJ235567
Psilotaceae[1]	*Psilotum nudum*	U30835	*Psilotum nudum*	U93822
Taxaceae	*Amentotaxus argotaenia*	L12580	*Taxus baccata*	AJ235619
Welwitschiaceae	*Welwitschia mirabilis*	AJ235814	*Gnetum Gnemon**	AF187060
Zamiaceae	*Zamia inermis*	L12683	*Zamia furfuracea*	AF188845

Appendix 1. *Continued*

	psaA		*PsbB*	
Family	**Species**	**G.B.#**	**Species**	**G.B.#**
Amborellaceae	*Amborella trichopoda*	AJ344262	*Amborella trichopoda*	AJ347847
Cupressaceae	*Sequoia sempervirens*	AF180012	*Sequoia sempervirens*	AJ347882
Cycadaceae	*Cycas taiwaniana*	AF180015	*Cycas taiwaniana*	AF222697
Ephedraceae	*Ephedra tweediana*	AF180017	*Ephedra tweediana*	AF222702
Ginkgoaceae	*Ginkgo biloba*	AF223226	*Ginkgo biloba*	AF222705
Nymphaeaceae	*Nymphaea sp.*	AJ344280	*Nymphaea sp.*	AJ347864
Pinaceae	*Pinus strobus*	AJ344296	*Pinus strobes*	AJ347880
Podocarpaceae	*Podocarpus gracilior*	AJ344297	*Podocarpus gracilior*	AJ347881
Psilotaceae[1]	*Psilotum nudum*	AF180023	*Psilotum nudum*	AF222707
Taxaceae	*Taxus brevifolia*	AJ344298	*Taxus brevifolia*	AJ347883
Welwitschiaceae	*Welwitschia mirabilis*	AF180013	*Welwitschia mirabilis*	AF222704
Zamiaceae	*Zamia pumila*	AJ344299	*Zamia pumila*	AJ347884

Appendix 1. *Continued*

	atpA		*cox1*	
Family	**Species**	**G.B.#**	**Species**	**G.B.#**
Amborellaceae	*Amborella trichopoda*	AF197711	*Amborella trichopoda*	AF193953
Cupressaceae	*Juniperus virginiana*	AF209106	*Juniperus virginiana*	AF020567
Cycadaceae	*Cycas revoluta*	AF20911	*Cycas revoluta*	AF020562
Ephedraceae	*Gnetum ula**	AF209109	*Ephedra viridis*	AF020564
Ginkgoaceae	*Ginkgo biloba*	AF209110	*Ginkgo biloba*	AF020565
Nymphaeaceae	*Nymphaea odorata*	AF209102	*Nymphaea odorata*	AF020570
Pinaceae	*Pinus strobus*	AF209108	*Pinus strobes*	AF020574
Podocarpaceae	*Podocarpus macrophyllus*	AF209105	*Podocarpus macrophyllus*	AF020575
Ophioglossaceae[1]	*Lycopodium digitatum**	AF209113	*Ophioglossum engelmannii*	AF020577
Taxaceae	*N/A*		*Taxus baccata*	AF020579
Welwitschiaceae	*Welwitschia mirabilis*	AF197618	*Welwitschia mirabilis*	AF020584
Zamiaceae	*Zamia furfuracea*	AF209111	*Zamia furfuracea*	AF020587

Appendix 1. *Continued*

	18S rRNA	
Family	**Species**	**G.B.#**
Amborellaceae	*Amborella trichopoda*	U42497
Cupressaceae	*Juniperus chinensis*	D38243
Cycadaceae	*Cycas taitungensis*	D85297
Ephedraceae	*Ephedra californica*	U42492
Ginkgoaceae	*Ginkgo biloba*	D16448
Nymphaeaceae	*Nymphaea sp.*	AF096696
Pinaceae	*Pinus elliottii*	D38245
Podocarpaceae	*Podocarpus costalis*	D38473
Psilotaceae[1]	*Psilotum nudum*	X81963
Taxaceae	*Taxus mairei*	D16445
Welwitschiaceae	*Welwitschia mirabilis*	AF207059
Zamiaceae	*Zamia pumila*	M20017

Note: G.B.# = GenBank accession number, * = Taxon substitution, 1= out group, N/A = Sequence not available

Appendix 2. Sequence of primers used in amplifying *rpa1* and *rpc1* and conditions of PCR

Both *rpa1* and *rpc1* were amplified in a single PCR reaction using degenerate primers. A universal primer for region D of the largest subunit of all RNA polymerases was used as forward primer. The sequence of this primer is: 5' CCI TAY AAY GCI GAY ITY GAY GGI GAY GAR ATG AA 3'. The reverse primer was designed from region G of aligned sequences of *rpc1* and *rpa1* of available taxa (mostly yeast species, plant sequences were not available at the time these primers were designed), using a computer software (CODEHOP, Rose et al. 1998) and manual inspection. The sequence of this primer is: 5'GCA AAA TGA AAA GTT TTC AGA GTC ATY TGN GT 3'

Gradient PCR (with annealing temperatures of 45, 50 and 55 °C) was done using 800-1000 ng of cDNA, 3 mM $MgCl_2$ and 2 units of Taq polymerase for each reaction. PCR was repeated for 35 cycles of denaturing (94 °C – 1 min), annealing (45 or 50 or 55 °C – 1 min) and extension (72 °C – 1 min). A final extension of 10 min at 72 °C was was also performed at the end of the 35 cycles. PCR products were separated on an agarose gel. In majority of cases two - three bands with the approximate size of 1.8 –2.0 kb were visible. These bands were excised from the gel (individually, when possible), purified using UltraClean15 kit (MOBIO Laboratories, Carlsbad, CA) and used in cloning.

Chapter 5

Comparative molecular evolution of the genes encoding the largest subunit of RNA ploymerase I, II and III

Abstract

Transcription in all cells is carried out by multi-subunit RNA polymerases. We studied the pattern and rate of molecular evolution of the genes encoding the largest subunit of the RNA polymerase I, II and III (*rpa1*, *rpb1* and *rpc1*, respectively) in seed plants. Our phylogenetic analysis clarifies the relationships of these genes and shows that that *rpb1* and *rpc1* share a more recent common ancestor relative to *rpa1*. The nonsynonymous substitution rate of *rpa1* and *rpc1* are approximately 4 and 2.5 times faster than that of *rpb1*. This rate is also lower in functional regions of these genes. Most of the amino acid positions are under high negative selection in *rpb1* whereas by *rpc1* and *rpa1* show increasingly more relaxed selective constraints. The pattern of functional constraint in these genes is generally in agreement with the known structural properties of the polypeptides they encode. Regions with important functions have higher functional constraint. Interestingly, the evolutionary rate of non-synonymous and translated amino acid sequences of *rpa1*, *rpb1* and *rpc1* corresponds to the level of concerted evolution in

the genes they transcribe. This suggests that the level of concerted evolution may have an impact on the evolution of the largest subunit of RNAPs.

Introduction

RNA polymerases (RNAPs) are complex multi-subunit enzymes responsible for synthesizing all cellular RNA molecules in the process of transcription. These enzymes share conserved regions from bacteria to human (Allison et al. 1985; Iwabe et al. 1991; Archambault and Friesen 1993). While bacteria and archaea have one RNAP, eukaryotes have three: RNA polymerase I (responsible for synthesizing ribosomal RNA precursors), RNA polymerase II (responsible for synthesizing pre-messenger RNA), and RNA polymerase III (responsible for synthesizing transfer RNAs and 5S ribosomal RNA). RNAP I, II and III have 8-14 subunits, the two largest subunits (subunits 1 and 2, subunit 1 being the largest and 2 the second largest) are homologous to the two largest subunits of *E.coli* RNAP (β' (Young 1991). There are also three other subunits with shared partial homologues in bacterial and archaeal RNAPs. These five subunits are called core subunits. The presence of these homologous subunits suggests a conservation of the function. In fact, the two largest subunits form the catalytic center of the enzyme in RNAPs. Bacterial RNAP (being the simplest RNAP compared to the eukaryotic RNAP) has been the focus of most of the studies on the structure and the function of RNAP. A major progress in understanding RNAP was reached in 1999 by determination of the crystallographic structure of the bacterial core RNAP (Zhang et al. 1999) and yeast RNAP II (Fu et al. 1999). By comparing the structures of the largest subunit of yeast RNAP II (RPB1) and bacterial (β' and β respectively), Ebright (2000) observed an even

higher degree of similarity than expected from primary sequence comparison. In addition, recent outstanding works of Kornberg's group provided a detailed analysis of the structural basis of transcription in yeast RNAP II (Cramer, Bushnell, and Kornberg 2001; Gnatt et al. 2001). The largest subunit of RNAP contains 8 evolutionary conserved blocks (known as regions A to H). Figure 1 shows a schematic diagram of the largest subunit of RNAP. The important functional regions, including the active site and regions of interaction with other subunits of the enzyme, are also shown on this diagram (this information is retrieved from Cramer, Bushnell, and Kornberg 2001). In this study we used the sequence information from region D to region G of the largest subunit of RNAP I, II and III in seed plants.

The genes encoding the largest subunit of RNA polymerase I, II and III (*rpa1*, *rpb1* and, *rpc1* respectively) evolved through duplications after the divergence of prokaryotes and eukaryotes (Iwabe et al. 1991). It is believed that *rpb1* and *rpc1* are evolutionarily closer (Memet, Saurin, and Sentenac 1988) and that *rpa1* evolves faster (Iwabe et al. 1991). However, due to the limited number of available *rpa1* and *rpc1* sequences most of the current knowledge on the molecular evolution of eukaryotic genes encoding the largest subunit of RNAPs was deduced from *rpb1* sequences. We cloned and sequenced ~ 1.8 kb of expressed copies of *rpa1*, *rpb1* and *rpc1* in all major groups of seed plants. This data set allowed us to study the molecular evolution of these genes in a time frame of approximately 300 million years (the approximate divergence time of seed plants). We show that there is a substantial difference between the evolutionary rate of *rpa1*, *rpb1* and *rpc1*.

We also studied the selective forces on individual amino acid sites of these three genes. It is widely known that the number of non-synonymous (amino acid changing) substitutions (Ka) per site divided by the number of synonymous (silent) substitutions (Ks) per site can reflect the evolutionary selective forces on a molecule (this ratio is called ω). A ratio of 1 indicates neutrality, a ratio >1 shows a probable positive selection and a ratio <1 shows purifying (negative) selection. Yang (Yang and Nielsen 2000; Yang et al. 2000) developed a maximum likelihood approach based on codon substitution models for estimating the posterior probability of ω categories for individual sites of an alignment. Using this approach, we show that a rather uniform and extremely negative selective pressure is imposed on *rpb1*. *rpa1* and *rpc1*, however, although mostly under negative selection, do not follow the same pattern and show regions with less or even relaxed functional constraints. We compared this information with known functional information derived from yeast *rpb1* structure (Cramer, Bushnell, and Kornberg 2001).

Materials and methods

Table 1 shows the list of genes and seed plant species used in this study. Cloning and PCR conditions were described elsewhere (see Chapters 2 and 4). Sequences of each gene were aligned at the amino acid level using ClustalW (Thompson, Higgins, and Gibson 1994). These alignments are shown in Appendix A. To build a phylogenetic tree of all *rpa1*, *rpb1* and *rpc1* genes, we also aligned all three genes together. Three bacterial RNAP (β' subunits) were used as out groups. This alignment is shown in Appendix A4. The nucleotide sequences were aligned based on the amino acid sequence. We used

MEGA2 software (Kumar et al. 2001) to perform G+C analysis, measure substitutions and to build phylogenetic trees. The number of synonymous (Ks) and non-synonymous (Ka) substitutions were calculated using the method of Pamilio-Bianchi-Li (Pamilo and Bianchi 1993). Amino acid distances (Kaa) were calculated according to a Poisson correction (Nei and Kumar 2000).

The PAML software (Yang 1997) was used to conduct site specific maximum likelihood analysis of ω. The likelihood model M8 (with 11 ω categories) was used for the analysis. This model resulted in the highest log likelihood for each data set and was chosen because it provides the largest number of ω categories. We also examined the pattern obtained from a simpler model (M3, with 3 ω categories) with no apparent difference (results not shown). Posterior probability values for ω categories were entered in a spread sheet and was color coded based on the variation obtained in each gene. We color coded different ranges of ω to be able to visually compare different levels of conservation/selection on each amino acid position in each gene. ω categories smaller than 0.1 were coded yellow, $0.11 \leq \omega \leq 0.2$ were coded blue, $0.21 \leq \omega \leq 0.35$ were coded green, $0.36 \leq \omega \leq 0.6$ were coded orange, and $\omega \geq 1.05$ were coded purple.

Results and discussion

A gene phylogeny of *rpa1*, *rpb1* and *rpc1* genes

Figure 2 shows a neighbor joining phylogenetic tree based on aligned *rpa1*, *rpb1* and *rpc1* translated amino acid sequences from 13 plant taxa and three bacterial RNAPs

(β'subunits) as outgroups. This tree shows that *rpa1*, *rpb1* and *rpc1* are all monophyletic. It also shows that *rpb1* and *rpc1* share a more recent common ancestor with one another than with *rpa1*. This branching order is supported by high bootstrap values and did not change as a result of changing the phylogenetic inference method or of using nucleotide sequences (results not shown). The branches leading to *rpb1* sequences are generally shorter than those leading to *rpa1* and *rpc1* sequences, suggesting that *rpb1* genes evolve more slowly than *rpa1* and *rpc1* genes. Contrary to what we obtained, a consensus branching order for these three genes does not exist in the literature. Memet et al. (1988) showed that these three genes share a common ancestor and that *rpb1* and *rpc1* are closer to one another than to *rpa1*. Iwabe et al. (1991) refuted this view and suggested that although these genes share a common ancestor, *rpa1* is closer to *rpb1* than either are to *rpc1*. A more recent study by Klenk et al. (1995) did not reach a consensus on the branching order of these genes, but generally supported the closeness of *rpb1* and *rpc1*. These studies, however, were all done with a small sampling of genes, specially *rpa1* and *rpc1* (for example, the largest data set included 4 *rpc1* and 5 *rpa1* sequences). Klenk et al. (1995) pointed out this problem and concluded that larger number of sequences would be needed to solve it. Our data set solves this problem and shows that the *rpb1* and *rpc1* genes are more closely related to one another than to *rpa1* genes.

G+C analysis of *rpa1*, *rpb1* and *rpc1*

The average %G+C content of *rpa1*, *rpb1* and *rpc1* are 41.6 (±1.0 SD) 42.1 (±0.7 SD) and, 41.8 (±0.7 SD), respectively. The average %G+C for each codon position is

shown in Figure 3. We plotted the %G+C at each codon position against %G+C of the gene, to see the effect of the G+C content of genes on the G+C content of codon positions. Since substitutions in the 1st and 2nd codon positions often lead to amino acid changes, a change in G+C content of the gene can influence the protein encoded by the gene. Our results show that this influence is minimal for *rpa1*, *rpb1* and *rpc1* (Figure 3). However, the increase in the G+C content of the gene increases the %G+C at the 3rd codon positions in *rpa1*, *rpb1* and *rpc1* (Figure 3).

Comparing rates of molecular evolution of *rpa1*, *rpb1* and *rpc1* genes

Table 1 shows that the length of the region D to G of *rpa1*, *rpb1* and *rpc1* shows different levels of conservation. The length of region D to G of *rpb1* genes is highly conserved in seed plants (average length of 1752.3 and SD of 2.5; only one gene is shorter by 9 nucleotides). The length of region D to G of *rpc1* genes is less conserved (average length of 1617.7 and SD of 40.2). For example, the length is identical in six gymnosperm species, but is 6 nucleotides shorter in the three closely related conifers species (*Taxus*, *Podocarpus* and *Thuja*). However, the length of region D to G of *rpc1* is different in 4 angiosperms. The length of region D to G of *rpa1* is not conserved in seed plants (average length of 1641.7 and SD of 61.1; the length is identical only in *Amborella* of angiosperms and *Ginkgo* of gymnosperms, which are not closely related). The difference in sizes of *rpc1* and *rpa1* genes is due to indels. The greater number of indels in these two genes is consistent with their faster rate of evolution (see below).

We computed all pairwise synonymous (Ks), non-synonymous (Ka) and amino acid (Kaa) distances for *rpa, rpb1* and *rpc1* genes (see Appendix B). Table 2 compares the mean evolutionary distances for each gene. Clearly, *rpb1* is the most conserved and slowest evolving (with a Ka of 0.06 and a Kaa of 0.1 substitutions per site), and *rpa1* is the most variable and fastest evolving (with a Ka of 0.24 and a Kaa of 0.38 substitutions per site). The difference between the Ka of these three genes is striking. *rpc1* has a Ka about 2.6 times, and *rpa1* has a Ka 4 times, higher than *rpb1*. A similar result is evident from comparing Kaa of these three genes. However, the synonymous substitutions of these genes do not show much difference (Table 2). Our results are in line with the only available study (Iwabe et al. 1991) that suggested a faster evolutionary rate for *rpa1*. However, Iwabe et al.'s (1991) data set included only the sequences of *rpa1, rpb1* and *rpc1* from yeast. The high level of conservation in *rpb1* genes and the lack of sequences of *rpa1* and *rpc1* from many species, has contributed to speculating a high evolutionary conservation for all RNA polymerases (Cramer, Bushnell, and Kornberg 2001). Our results shows that this generalization is not appropriate when comparing RNAP I, II and III between eukaryotes from a relatively large phylogenetic range.

To compare the evolutionary rate of different parts of region D to G of *rpa1, rpb1* and *rpc1*, we estimated Ks, Ka and Kaa of each conserved region and the regions in between them (Table 3; Figure 4). Ks values do not show a large variation between different regions and different genes (Figure 4A). This suggests that the rate of synonymous substitutions is relatively uniform in different parts of these genes. However, Ka and Kaa values, while in agreement with each other, show a large variation in different regions (Figure 4B and C, respectively). Region F and region F-G have the

lowest and the highest rate of non-synonymous and amino acid substitutions per site in all three genes, respectively. Interestingly, the degree of nonsynonymous and amino acid variation is not the same between different regions of these genes. For example while the Ka value of the region F of *rpa1* is 3 times larger than the Ka value of the same region in *rpb1*, the Ka value of the region D-E of *rpa1* is 13 times larger than the same region in *rpb1* (Figure 4B). Interestingly, region D-E in *rpa1* has an insertion of about 17 amino acids that does not exist in the other two genes.

Studying selective forces on each site

Figures 5, 6 and 7 show the results of maximum likelihood analysis of ω for each amino acid position for *rpa1*, *rpb1* and *rpc1* respectively. Figure 5 shows that *rpb1* is more or less uniform under strong negative selection. A majority of the amino acid positions of this gene have $\omega \leq 0.08$ and positions with higher ω belong to the $\omega \leq 0.189$ category (Figure 5). There is only one position which belongs to the $\omega = 1.064$ category, suggesting that this site is evolving neutrally. A different pattern is observed in *rpa1* and *rpc1* sequences. Figure 6 shows the result obtained for *rpc1*. *rpc1* is under negative selection in most of its amino acid positions (yellow color in Figure 6), but there are also sites which belong to higher ω categories. There are also several amino acid positions which belong to the $\omega = 1.058$ category. Sites with lower functional constraints are in fact far from region D, which is known to contain the active site of the RNAPs. The largest number of sites with low/relaxed negative selection is close to the region G of the enzyme. Figure 7 shows that *rpa1* is even under a lower level of negative selection.

Although there are regions with a high degree of negative selection, there are also many sites that belong to higher ω categories (i.e., ω=0.584; orange color in Figure 7). There is also one site which belongs to the ω=14.423 category, suggesting a probable positive selection at this position. This site is in the position 364 of the alignment of *rpa1* (See Appendix A). It is located in between the conserved regions E and F. In yeast RPB1, this region is shown to be involved in the interaction with the subunit RPB9 of the enzyme (see Figure 1). This site contains a variety of different amino acids in different plant species (it is only the same in three species: *Zea* and *Oryza* from angiosperms and *Ephedra* from gymnosperms). However, based on the Taylor's physiochemical classification (Taylor 1986), the amino acids at this position mostly belong to the polar class (except glycine in *Zamia* and cysteine in *Pinus* which are classified as tiny and small amino acids respectively).

Linking evolutionary constraint to structural properties

Figure 8 compares the pattern of evolutionary functional constraint in *rpa1*, *rpb1* and *rpc1*. It also shows the map of the well-conserved regions D to G. Here we compare this evolutionary functional constraint with what is known about the structural properties of the region D to G of the RNAPs (Cramer, Bushnell, and Kornberg 2001). Region D contains the active site of the enzyme, i.e., the three metal binding aspartate residues (invariant in all RNAPs, Cramer 2002). We only obtained the sequence of the active site in *rpb1*, and in fact in the other two genes we used this invariant part to amplify the genes (it was part of the primer used for PCR and is therefore not included in our alignment of

rpa1 and *rpc1*). Region D-E is not functionally important in RPB1 of yeast (Cramer, Bushnell, and Kornberg 2001). This region is functionally less constrained in RPA1 and RPC1 than RPB1. Region E is important for interaction with subunit RPB8 of the enzyme (Figure1). This subunit is common between the three RNAP I, II and III (they all use the same subunit, Cramer 2002). Interestingly sequence elements of region E are under strong functional constraint in *rpa1, rpb1* and *rpc1* (Figure 8, compare the middle part of the region E in the three profiles). The region in between conserved regions E and F is involved in interaction with subunits RPB2 and RPB9 (Figure 1 and, Cramer, Bushnell, and Kornberg 2001). These two subunits are not common between the three eukaryotic RNAPs (Cramer 2002). RPB2 is among core subunits (homologues in all RNAPs) and its homologues in RNAP I and III are known (Cramer 2002). The region of interaction with RPB2 is under strong functional constraint in the *rpb1* profile. The same part in the *rpc1* and *rpa1* profiles is under high functional constraint, suggesting that this region, just like in *rpb1*, might be functionally important in interacting with other subunits of RNAP I and III. The region of interaction of RPB1 with RPB9 (approximately between residues 200 – 260 on the alignment, see Appendix A4 and Figure 1), however, is under less functional constraint in *rpa1* and *rpc1*. Interestingly, RPB9 is neither a common nor a core subunit. It was suggested that RPB9 has an important role in straitening the DNA (before binding to the enzyme and melting to make a transcription bubble) in the initiation phase of the transcription (Cramer, Bushnell, and Kornberg 2001). Since the segment of interaction with RPB9 (which is not located in the homologues regions A to H of all RNAPs) is under a high degree of functional constraint in *rpb1* and the same region in *rpc1* and *rpa1* is not under the same high functional

constraint, there might be different structural/functional properties for this region in *rpc1* and *rpa1*.

Region F is the largest conserved region in the largest subunit of RNAPs. This region is important because it forms a key functional part of the enzyme called the bridge. In the 3D structure of the RNAP, bridge is close to the active site and is responsible for ordering the DNA template strand, by binding to it, before the active site starts transcribing (Cramer, Bushnell, and Kornberg 2001; Cramer 2002). In fact, this region is under a high degree of functional constraint in *rpa1, rpb1* and *rpc1* (Figure 8, compare block F in three panels). The segment between conserved regions F and G is not considered to be a functionally important part of the RNAP structure and the pattern of functional constraint in Figure 8 agrees with this (compare region between regions F and G in three panels).

In conclusion, the pattern of evolutionary functional constraint in the region D to G of *rpa1, rpb1* and *rpc1* generally agrees with known structural properties of this part of RNAP. However, regions in between the conserved regions are functionally less constrained. One region with known functional properties in RNAP II, is not functionally constrained in RNAP I and III (the region between regions E and F), suggesting a probable difference in the function/structure of this region in RNAP I and III when compared with RNAP II.

Putting it all together

The evolutionary functional constraint clearly follows known structural properties of the RNAPs. However, there remains an important question: why do these three enzymes (with the same functional domains) evolve with three different rates? Our analysis of 300 million years of evolution of these molecules in seed plants, with a uniform sampling (the same species were used to avoid organism specific molecular evolution), allowed us to clearly show that *rpa1* and *rpc1* are approximately evolving 4 and 2.6 times faster (at non-synonymous sites of the D-G region; see Table 2) than *rpb1*, respectively. Although we see a strong negative selection in the most important functional parts of these genes, *rpa1* and *rpc1* are functionally less constrained in comparison with *rpb1*.

We think there might be a link with the type of genes that each RNAP transcribes and their rate and pattern of molecular evolution. RNAP II is responsible for mRNA synthesis of protein-coding genes, whereas RNAP I transcribes ribosomal RNA genes and RNAP III transcribes tRNA and 5S ribosomal RNA genes. It is widely known that rRNA genes undergo frequent exchanges (e.g., through unequal crossing over and gene conversion), which leads to their concerted evolution (Dover 1982; Arnheim 1984). These mechanisms therefore lead to the fixation of species specific substitutions in the promoter region of genes evolving in concert. This phenomenon could therefore explain why a human protein-coding gene can be readily transcribed using a yeast promoter, whereas, an rRNA gene from one species cannot be transcribed in another species (e.g.,

from mouse to human). Dover and Flavell (1984) suggested that RNAP I has to adapt to the accelerated rate of evolution of rRNA promoter brought about by their concerted evolution. 5S rRNA and tRNA genes are also known to evolve in a concerted fashion (Amstutz et al. 1985; Drouin and Moniz de Sa 1995) but likely more slowly than rRNA genes because they are often found dispersed throughout the genome whereas rRNA genes are most often found in a few cluster of tandemly repeated genes containing dozens to hundreds of genes. In contrast, the promoters of protein coding genes (transcribed by RNAP II) will not evolve fast due to concerted evolution because they are found upstream of different genes. The proteins involved in transcribing rRNA and tRNA genes, are therefore expected to evolve faster than those involved in transcribing protein coding genes. For example Heix and Grummt (1995) and Heix et al. (Heix et al. 1997) have shown that the transcription factors of RNAP I, are species-specific (Heix and Grummt 1995; Heix et al. 1997). They suggest that subtle differences in the interaction of transcription factors with species-specific elements and components of the transcription apparatus are the major factors in the species specificity of rDNA transcription (Heix and Grummt 1995). Our results show that the rate of evolution of the largest subunit of RNAP I, II and III is correlated with the amount of concerted evolution in the genes they transcribe. In other words, *rpa1* and *rpc1* (with a lower degree) evolve (change) more rapidly than *rpb1* to adapt to the concerted changes in the promoter regions of the genes they transcribe.

While our data set allowed us to compare a functionally important part of the largest subunit of RNAPs in several taxa, it lacks sequence information from other important parts of the RNAPs complex molecule. For example, region A in the largest

subunit of RNAP II is known to be involved in interaction with transcription factors (Cramer et al. 2001, 2002) and it would be interesting to compare its evolution in RNAP I, II and III.

References

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

Amstutz, H., P. Munz, W.-D. Heyer, U. Leupold, and J. Kohli. 1985. Concerted evolution of tRNA genes: intergenic conversion amnog three unlinked serine tRNA genes in S. pombe. Cell **40**:879-886.

Archambault, J., and J. D. Friesen. 1993. Genetics of eukaryotic RNA polymerases I, II, and III. Microbiol. Rev. **57**:703-724.

Arnheim, N. 1984. Concerted evolution of multigene families. Pp. 38-61 *in* R. Koehn, ed. Evolution of genes and proteins. Sinauer Inc., Sunderland, MA.

Cramer, P. 2002. Multisubunit RNA polymerases. Curr. Opin. Struc. Biol. **12**:89–97.

Cramer, P., D. Bushnell, and R. Kornberg. 2001. Structural basis of transcription: RNA polymerase II at 2.8 Å resolution. Science **292**:1863-1876.

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

Dover, G., and R. B. Flavell. 1984. Molecular coevolution: DNA divergence and maintenance of function. Cell **38**:622-623.

Drouin, G., and M. Moniz de Sa. 1995. The concerted evolution of 5S ribosomal genes linked to the repeat units of other multi gene families. Mol. Biol. Evol. **12**:481-493.

Ebright, R. 2000. RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. J. Mol. Biol. **304**:687-698.

Fu, J., A. Gnatt, D. Bushnell, G. Jensen, N. Thompson, R. Burgess, P. David, and R. Kornberg. 1999. Yeast RNA polymerase II at 5 Å resolution. Cell **98**:799-810.

Gnatt, A., P. Cramer, J. Fu, D. Bushnell, and R. Kornberg. 2001. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Å resolution. Science **292**:1876-1882.

Heix, J., and I. Grummt. 1995. Species specificity of transcription by RNA polymerase I. Curr. Opin. Genet. Dev. **5**:652-656.

Heix, J., J. C. Zomerdijk, A. Ravanpay, R. Tjian, and I. Grummt. 1997. Cloning of murine RNA polymerase I-specific TAF factors: conserved interactions between the subunits of the species-specific transcription initiation factor TIF-IB/SL1. Proc. Natl. Acad. Sci. USA **94**:1733-1738.

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

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

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17**:1244-1245.

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

Nei, M., and K. Kumar. 2000. Molecular evolution and phylogenetics. Oxford university press.

Pamilo, P., and N. Bianchi. 1993. Evolution of the Zfx and Zfy genes : rates and interdependence between the genes. Mol. Biol. Evol. **10**:271-281.

Taylor, W. R. 1986. The Classification of Amino Acid Conservation. J. Theor. Biol. **119**:205 - 218.

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

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555-556.

Yang, Z., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17**:32-43.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431-449.

Young, R. 1991. RNA polymerase II. Ann. Rev. Biochem. **60**:689-715.

Zhang, G., E. Campbell, L. Minakhin, C. Richter, K. Severinov, and S. Darst. 1999. Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 Å resolution. Cell **98**:811-824.

Table 1. Names and length (in nucleotides) of region D-G of the *rpa1*, *rpb1* and *rpc1* genes used in this study

Plant group	Species	*rpa1*	*rpb1*	*rpc1*
Angiosperm	*Amborella trichopoda*	AmbA1(1605)	AmbB1(1753)	N/A
Angiosperm	*Arabidopsis thaliana*	AraA1(1668)	AraB1(1753)	AraC1(1584)
Angiosperm	*Drimys winteri*	DriA1(1602)	N/A	DriC1(1614)
Angiosperm	*Nymphaea odorata*	N/A	NymB1(1753)	NymC1(1602)
Angiosperm	*Oryza sativa*	OryA1(1794)	OryB1(1753)	OryC1(1611)
Angiosperm	*Zea mays*	ZeaA1(1743)	N/A	N/A
Gymnosperm	*Cycas revoluta*	CycA1(1620)	CycB1(1753)	CycC1(1611)
Gymnosperm	*Ephedra viridis*	EphA1(1581)	EphB1(1753)	EphC1(1611)
Gymnosperm	*Ginkgo biloba*	GinA1(1605)	GinB1(1753)	GinC1(1611)
Gymnosperm	*Pinus nigra*	PinA1(1641)	PinB1(1753)	PinC1(1611)
Gymnosperm	*Podocarpus macrophyllus*	PodA1(1629)	PodB1(1753)	PodC1(1605)
Gymnosperm	*Taxus canadensis*	TaxA1(1614)	TaxB1(1753)	TaxC1(1605)
Gymnosperm	*Thuja occidentalis*	N/A	ThuB1(1753)	ThuC1(1605)
Gymnosperm	*Welwitschia mirabilis*	WelA1(1608)	WelB1(1753)	WelC1(1611)
Gymnosperm	*Zamia muricata*	ZamA1(1632)	ZamB1(1744)	ZamC1(1611)

Table 2. Nucleotide and amino acid variation and mean synonymous (Ks), non-synonymous (Ka) and amino acid (Kaa) distances.

Gene	%nuc con	%nuc var	%aa con	%aa var	Mean Ks ± SE	Mean Ka ± SE	Mean Kaa ±SE
rpa1	34	66	38.1	61.9	1.430 ± 0.050	0.244 ± 0.012	0.383 ± 0.021
rpb1	55	45	71.6	28.4	1.277 ± 0.048	0.061 ± 0.005	0.101 ± 0.008
rpc1	40.2	59.8	46.2	53.8	1.323 ± 0.050	0.159 ± 0.010	0.256 ± 0.015

Notes: nuc= nucleotoide; aa= amino acid, con= conserved; var= variable; SE= standard error calculated by bootstrapping (100 replicates)

Table 3. Positions of conserved regions E, F and G and regions in between them (D-E, E-F, F-G) on the alignment of translated amino acid sequences of *rpa1, rpb1* and *rpc1*. See appendix A for the alignments. Numbers for each region correspond to starting and ending amino acids.

	D-E	E	E-F	F	F-G	G
rpa1	7 - 92	93 - 130	131 - 420	421 - 528	529 - 640	641 - 664
rpb1	7 - 69	70 - 108	109 - 257	258 - 365	366 - 558	559 - 585
rpc1	7 - 69	70 - 109	110 - 271	272 - 409	410 - 547	548 - 584

Figure 1. A schematic diagram of the largest subunit of RNAPs. Conserved regions A-H, active site and important interacting regions are shown.

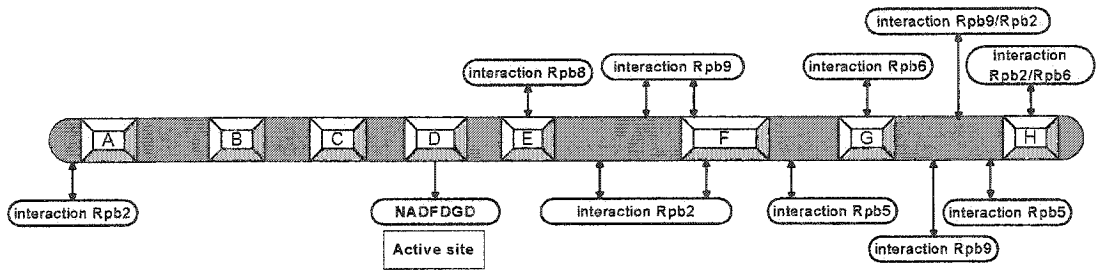


Figure 2. Phyloegentic tree of *rpa1*, *rpb1* and *rpc1* genes in seed plants and three bacterial β' subunits as outgroups. The tree was reconstructed from translated amino acid sequences using the neighbor joining method (with distances calculated using a Poisson correction). Bootstrap values are shown on major branches of the tree.

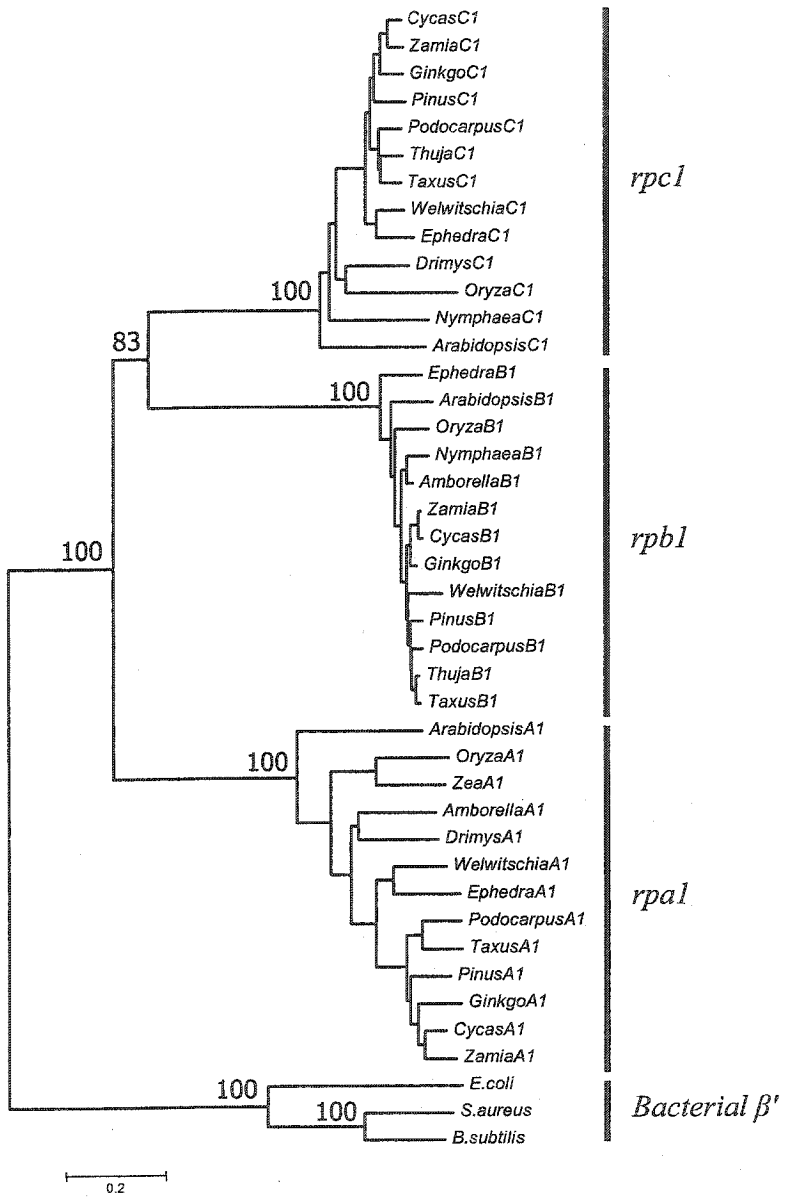
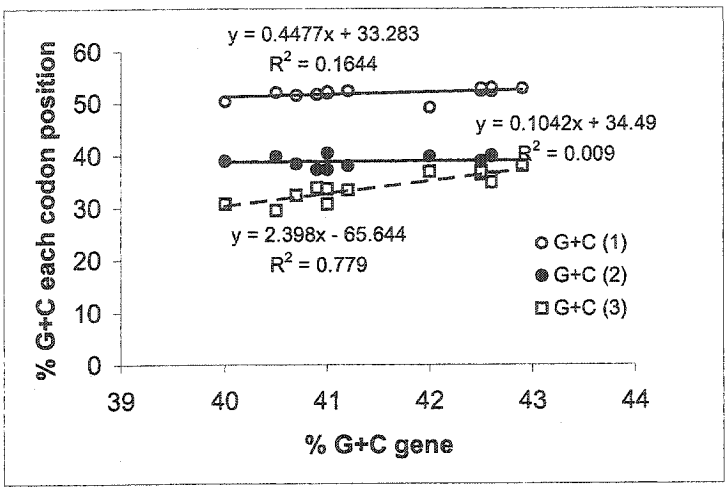
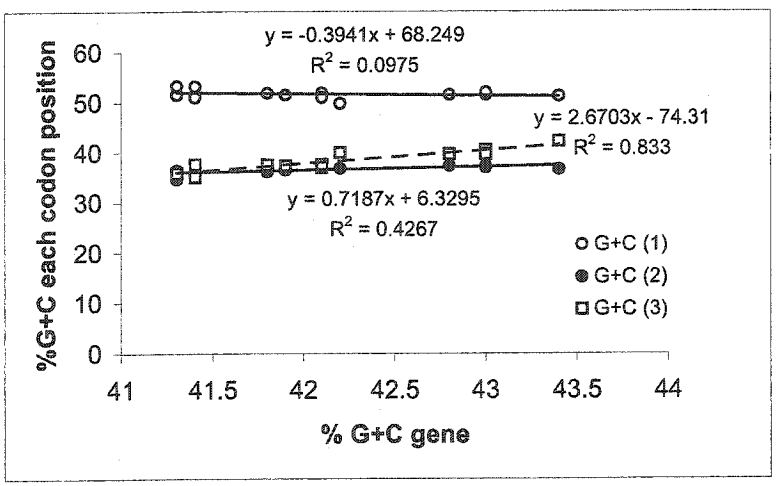
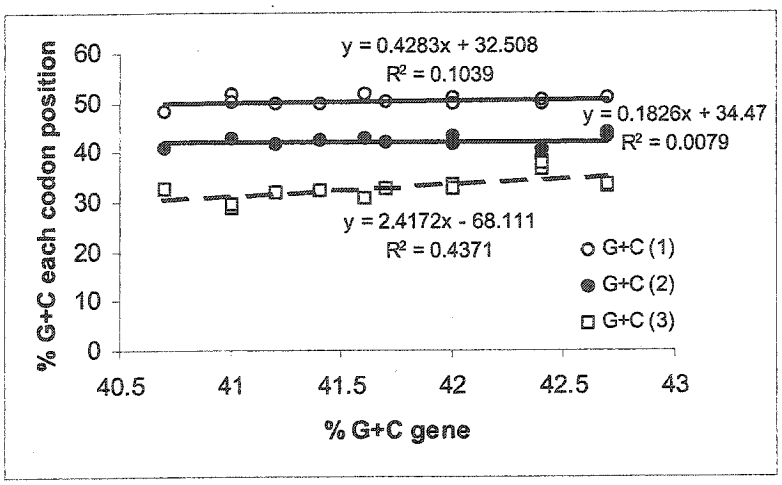


Figure3. (A) G+C analysis at each codon position. (B) Average %G+C at each codon positions ± standard error.

A







B

rpa1

Avg 1st 51.9±0.3

Avg 2nd 38.3±0.3

Avg 3rd 34.0±0.7

rpb1

Avg 1st 51.6±0.3

Avg 2nd 36.6±0.2

Avg 3rd 38.2±0.6

rpc1

Avg 1st 50.4±0.3

Avg 2nd 42.1±0.4

Avg 3rd 32.8±0.7

Figure 4. Comparison of the mean number of amino acid (Kaa), non-synonymous (Ka) and synonymous (Ks) substitutions in conserved regions E, F and G and regions in between (D-E, E-F, F-G). Note that region G is partial. Error bars represent standard errors calculated using bootstrapping (100 replicates).

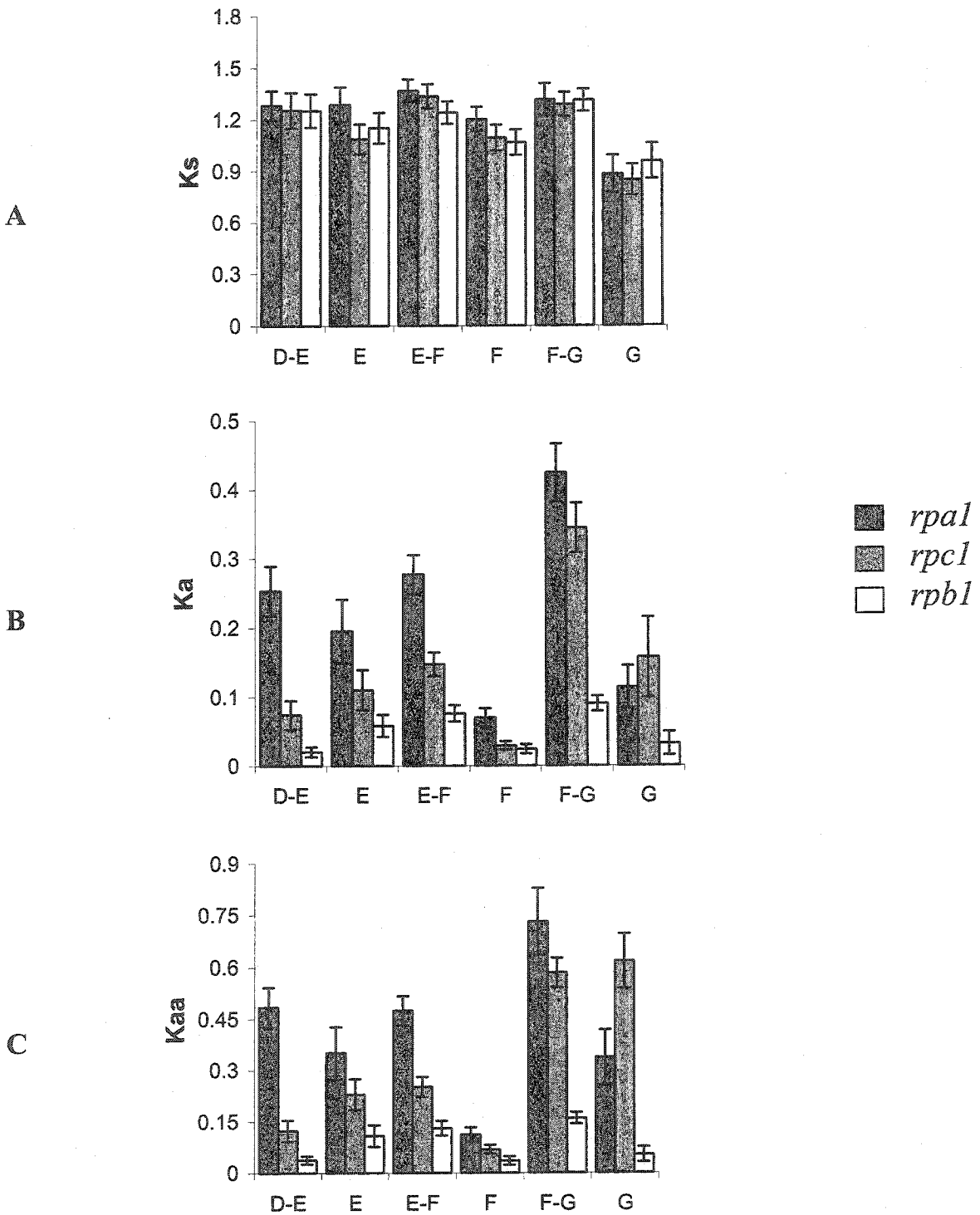


Figure 5. Site-specific pattern of evolutionary functional constraint for region D-G of *rpb1*. Indels are removed from the analysis.

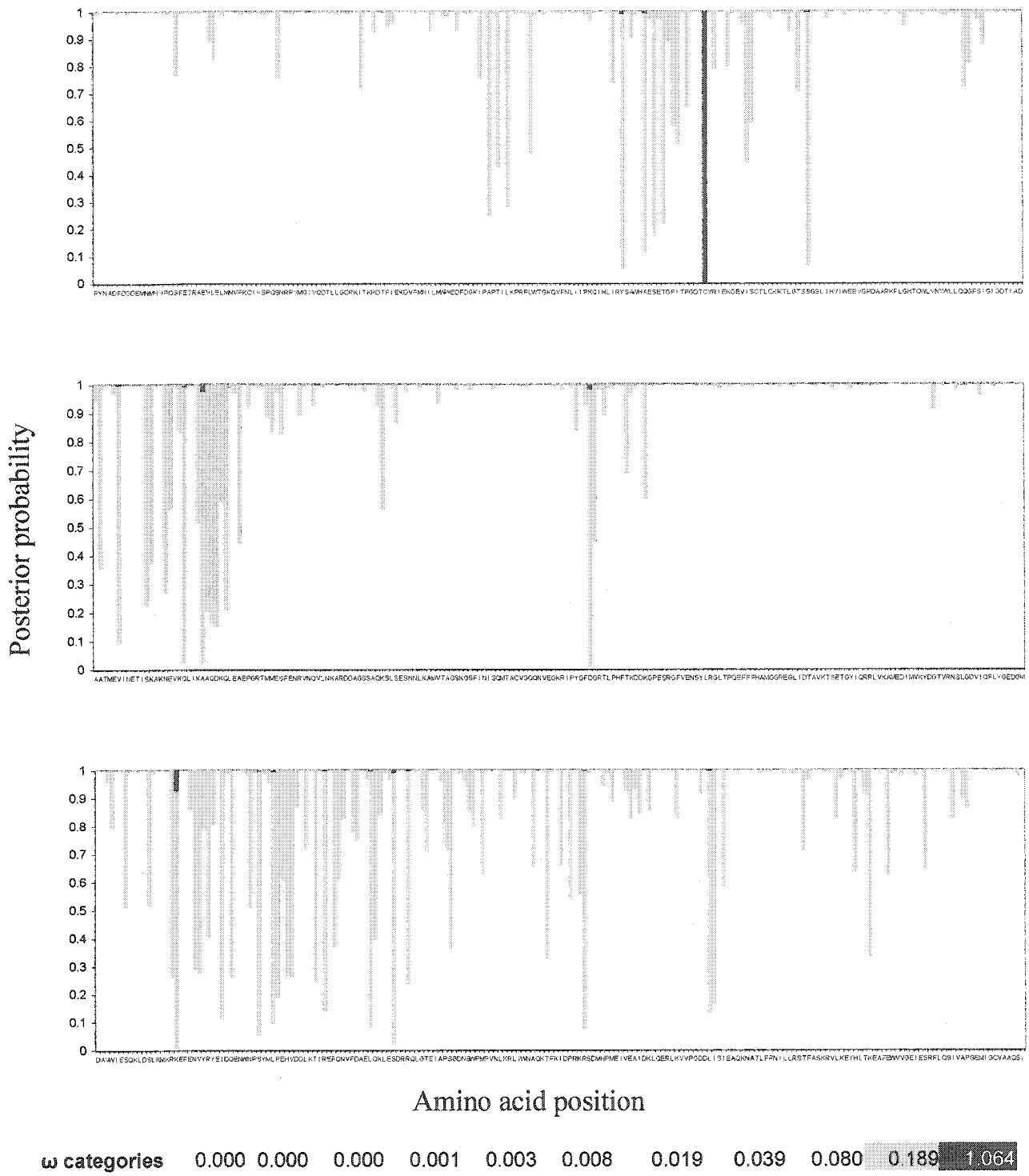


Figure 6. Site-specific pattern of evolutionary functional constraint for region D-G of *rpc1*. Indels are removed from the analysis.

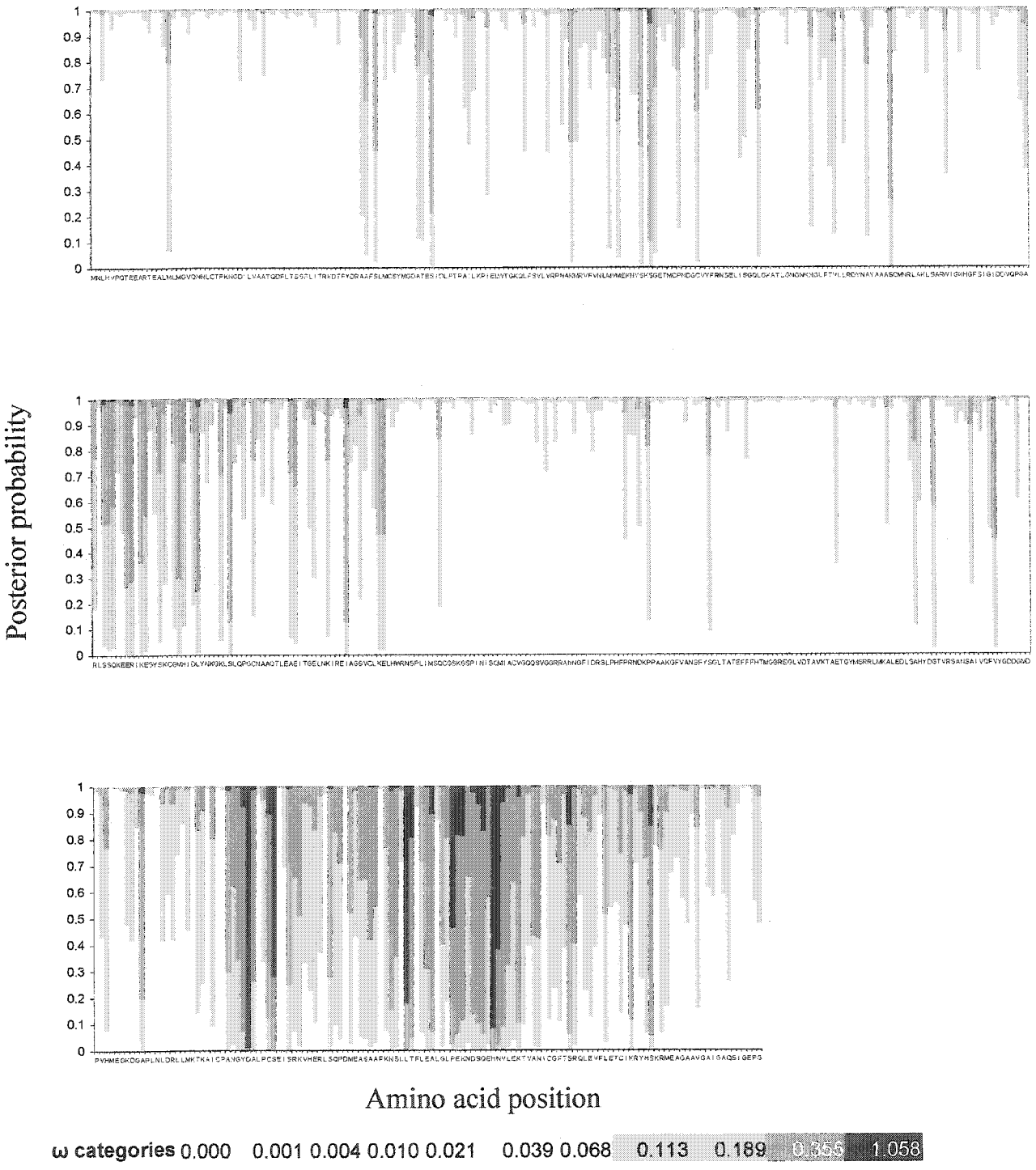


Figure 7. Site-specific pattern of evolutionary functional constraint for region D-G of *rpa1*. Indels are removed from the analysis.

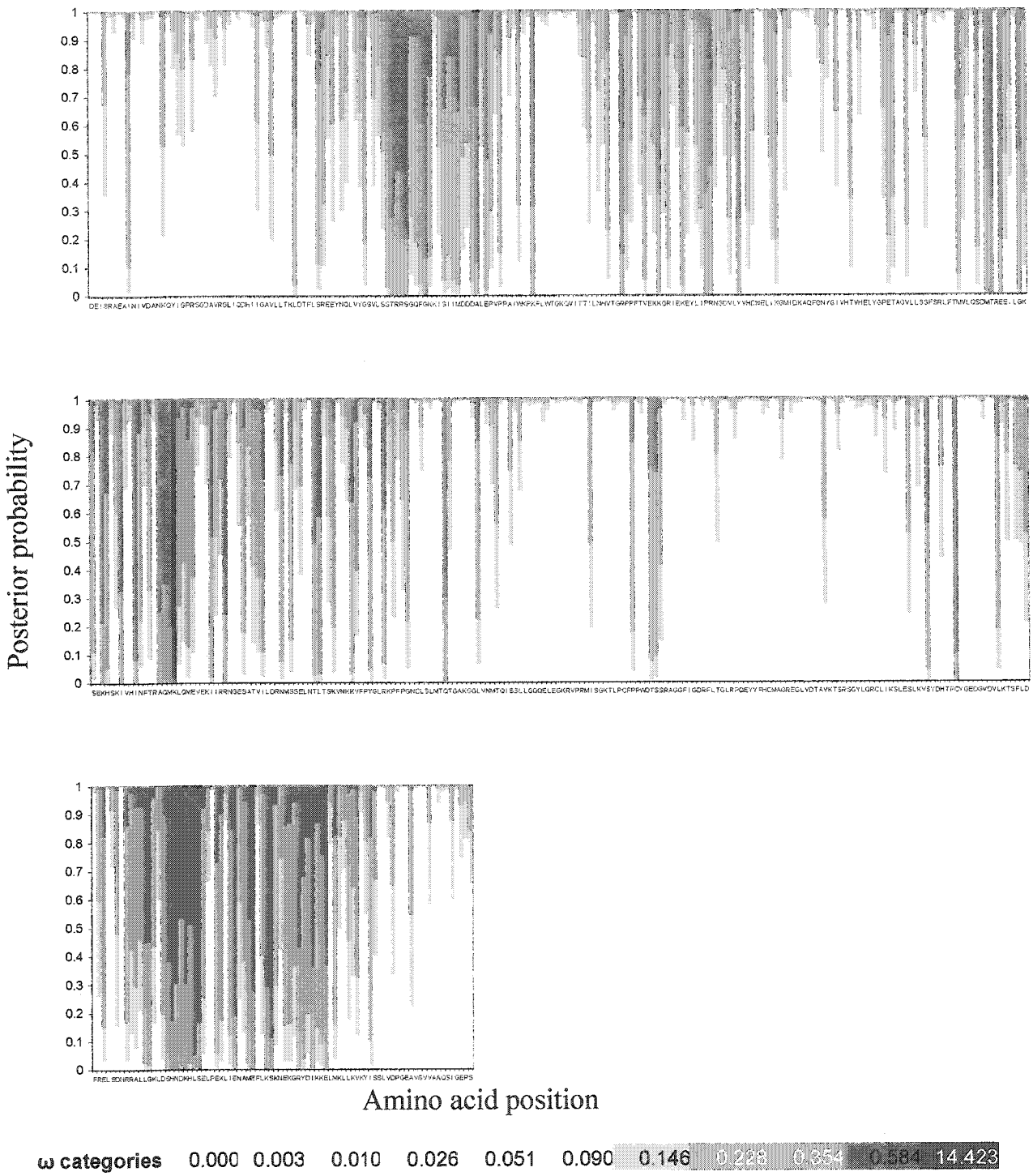
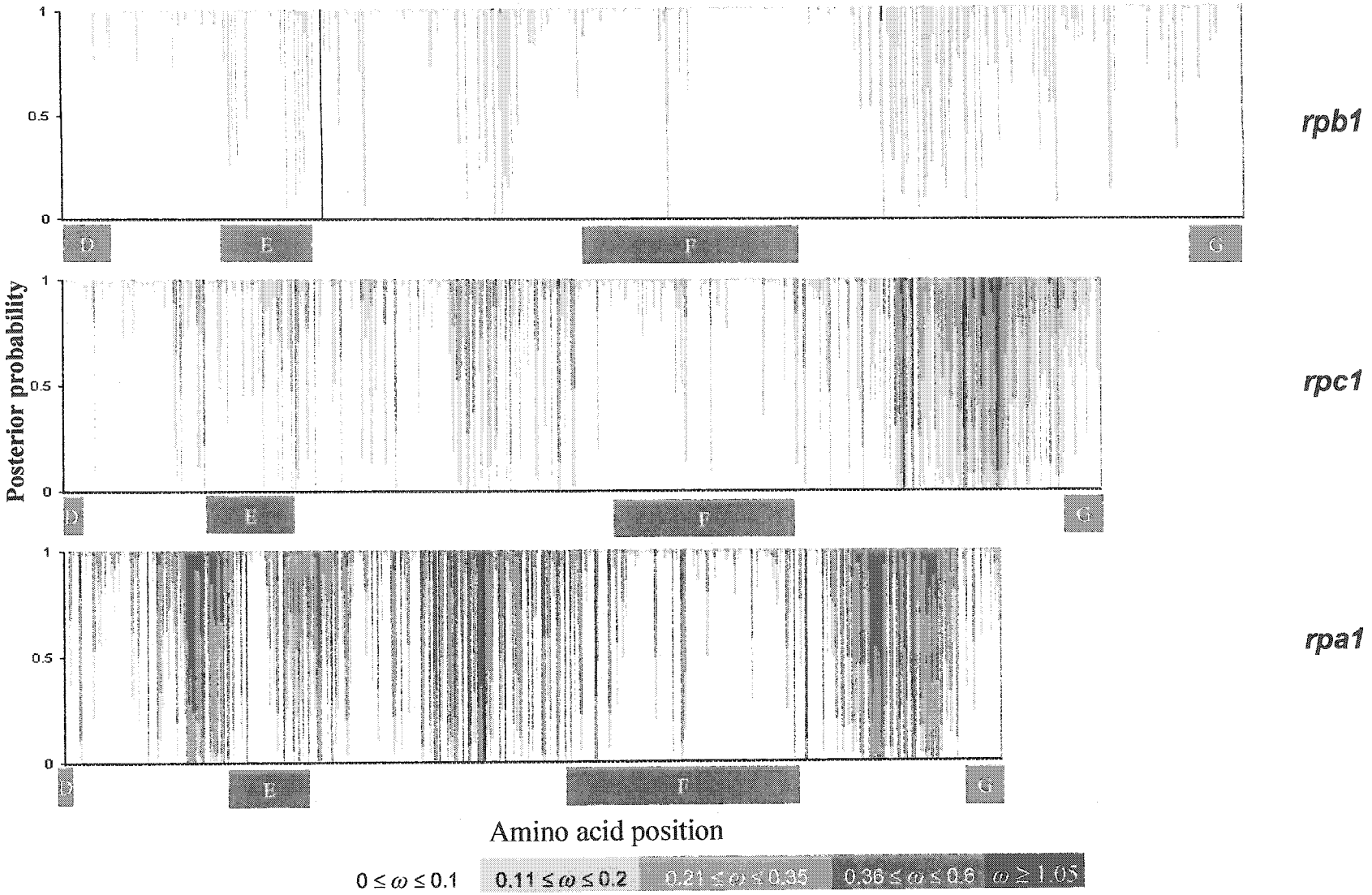
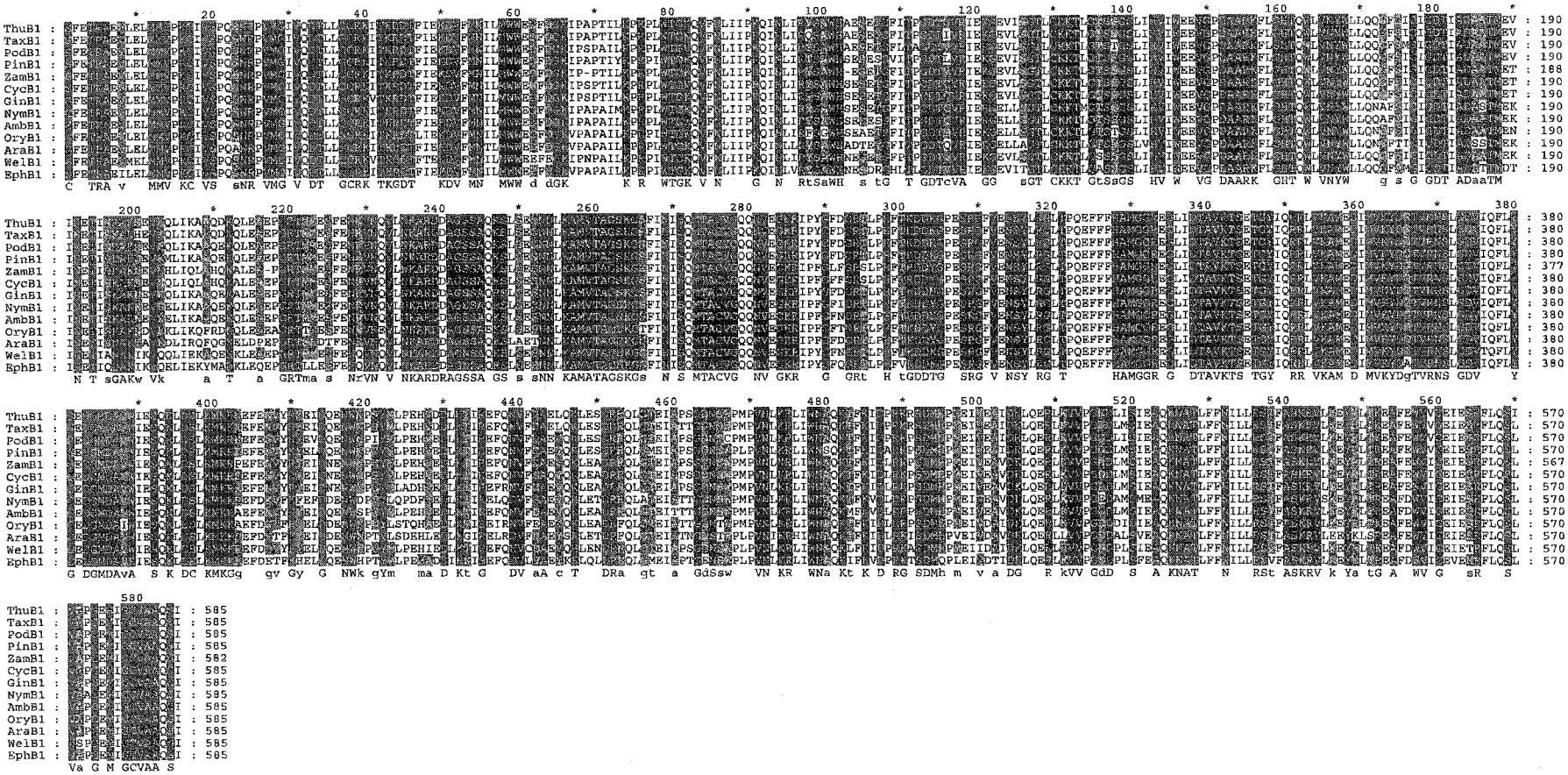


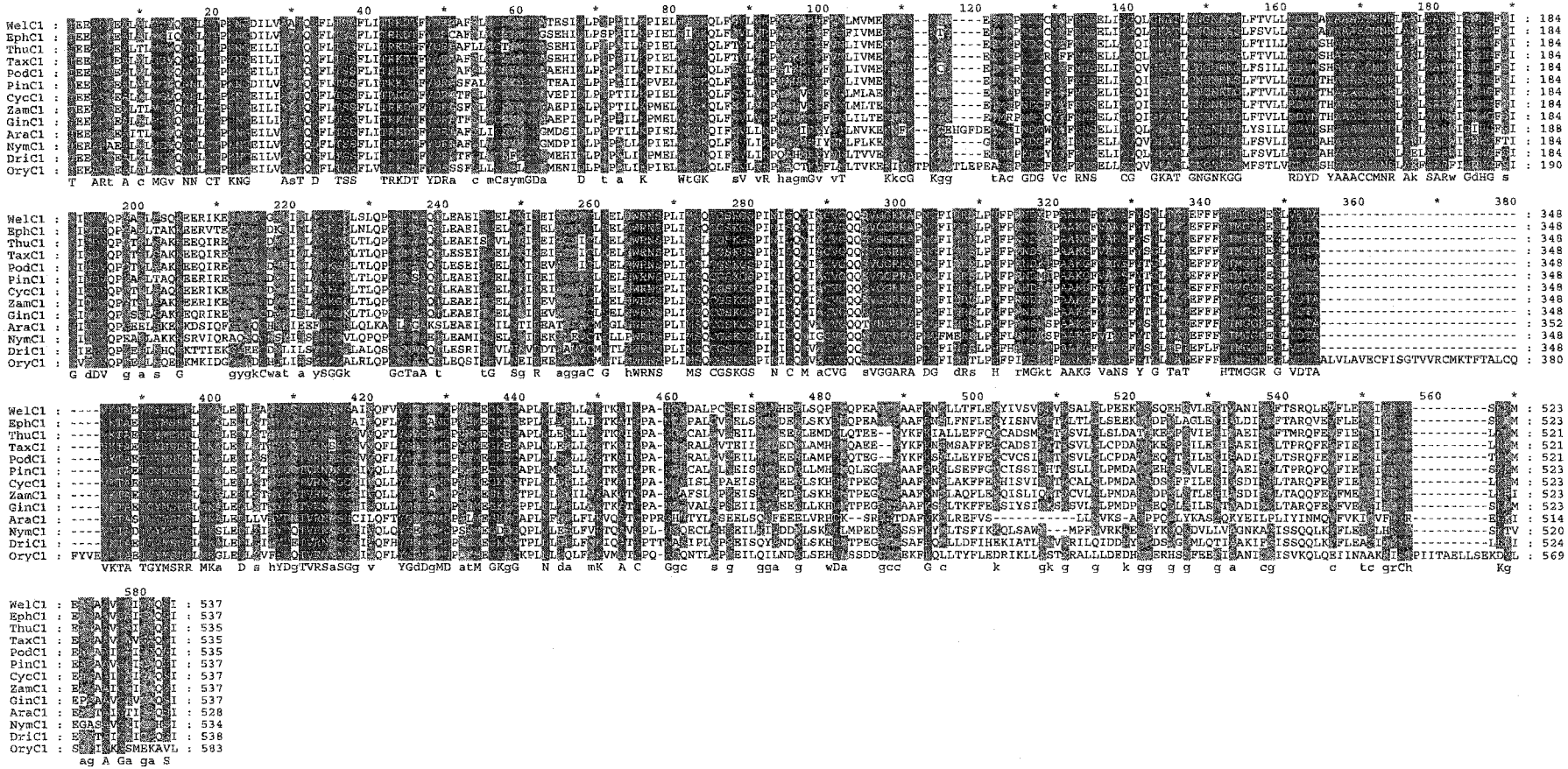
Figure 8. Comparison of the site-specific pattern of functional constraint in *rpa1*, *rpb1* and *rpc1*. Regions D-G are shown on the alignments. Sizes of the pictures are approximately scaled to the size of each gene. ω color scheme is approximate, for the detailed ω range for each gene see pictures 2, 3 and 4. Note that conserved regions D and G were used in the amplification of the genes and are partial in each gene. Sizes of the conserved blocks are proportional to the alignment of each gene. Indels are removed from the analysis.



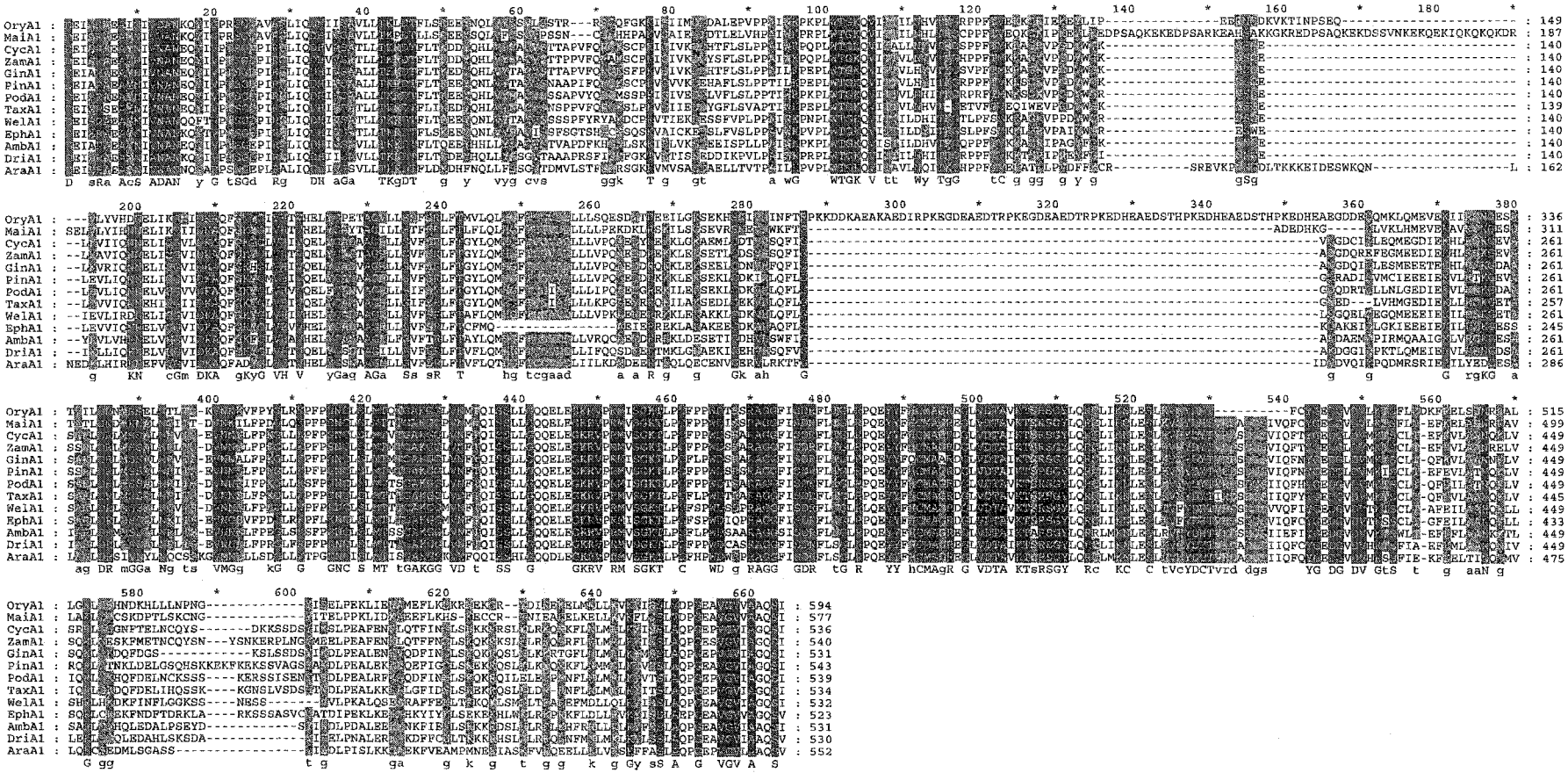
Appendix A1. Alignment of the translated amino acid sequences of *rpb1* genes. The alignment is shaded for three levels of conservation. See Table 1 for abbreviations.



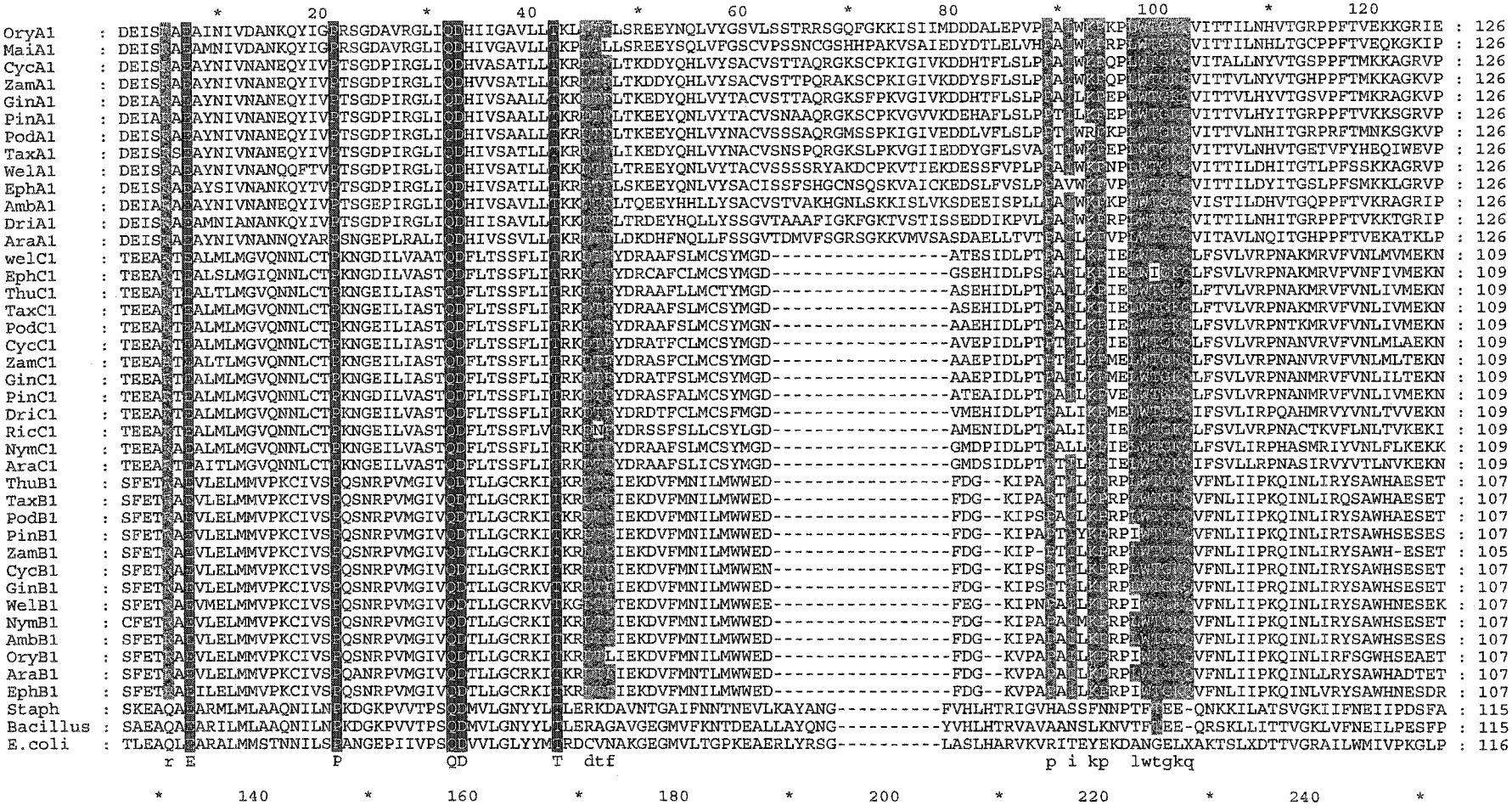
Appendix A2. Alignment of the translated amino acid sequences of *rpc1* genes. The alignment is shaded for three levels of conservation. See Table 1 for abbreviations.



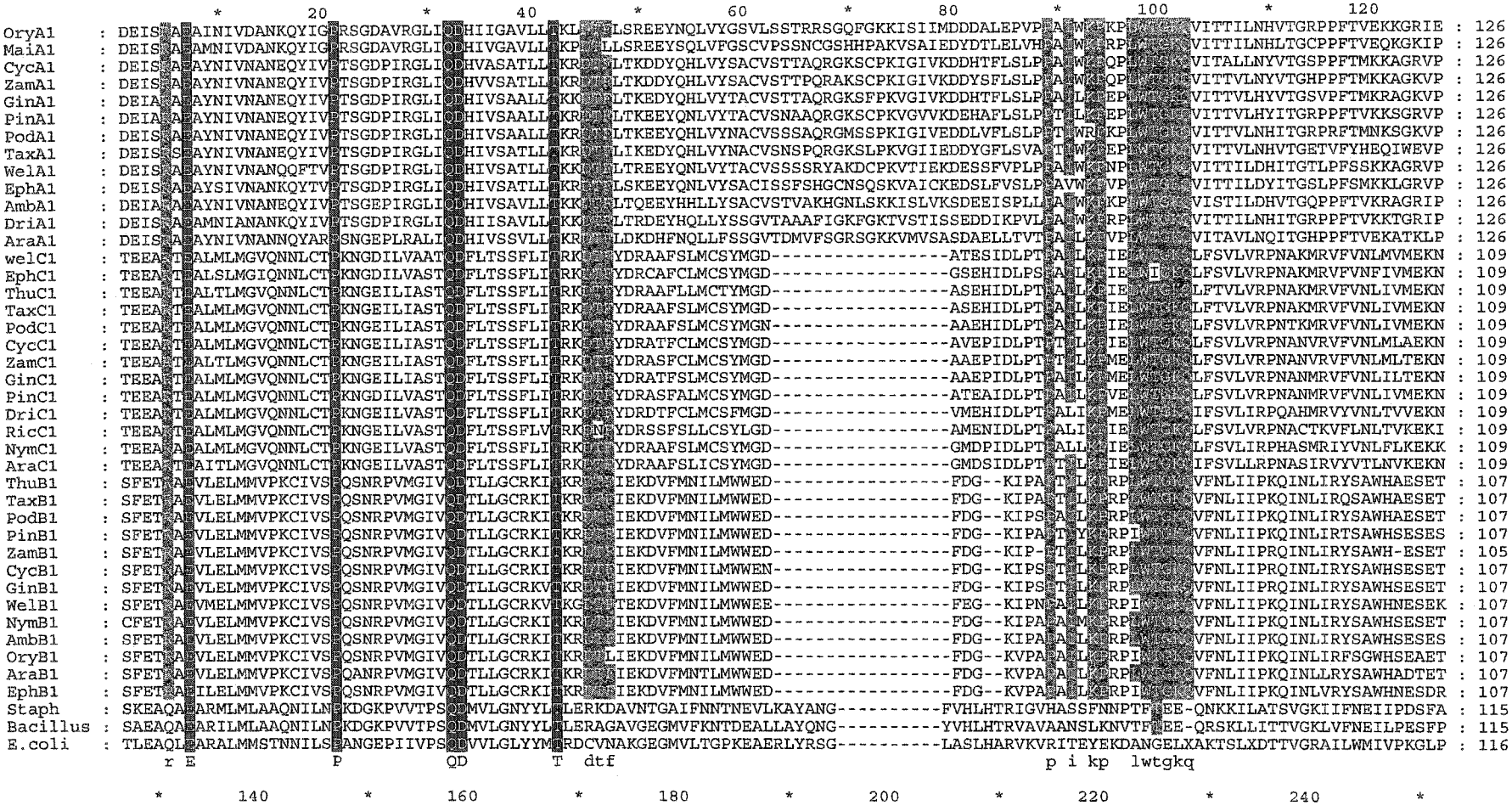
Appendix A3. Alignment of the translated amino acid sequences of *rpa1* genes. The alignment is shaded for three levels of conservation. See Table 1 for abbreviations.



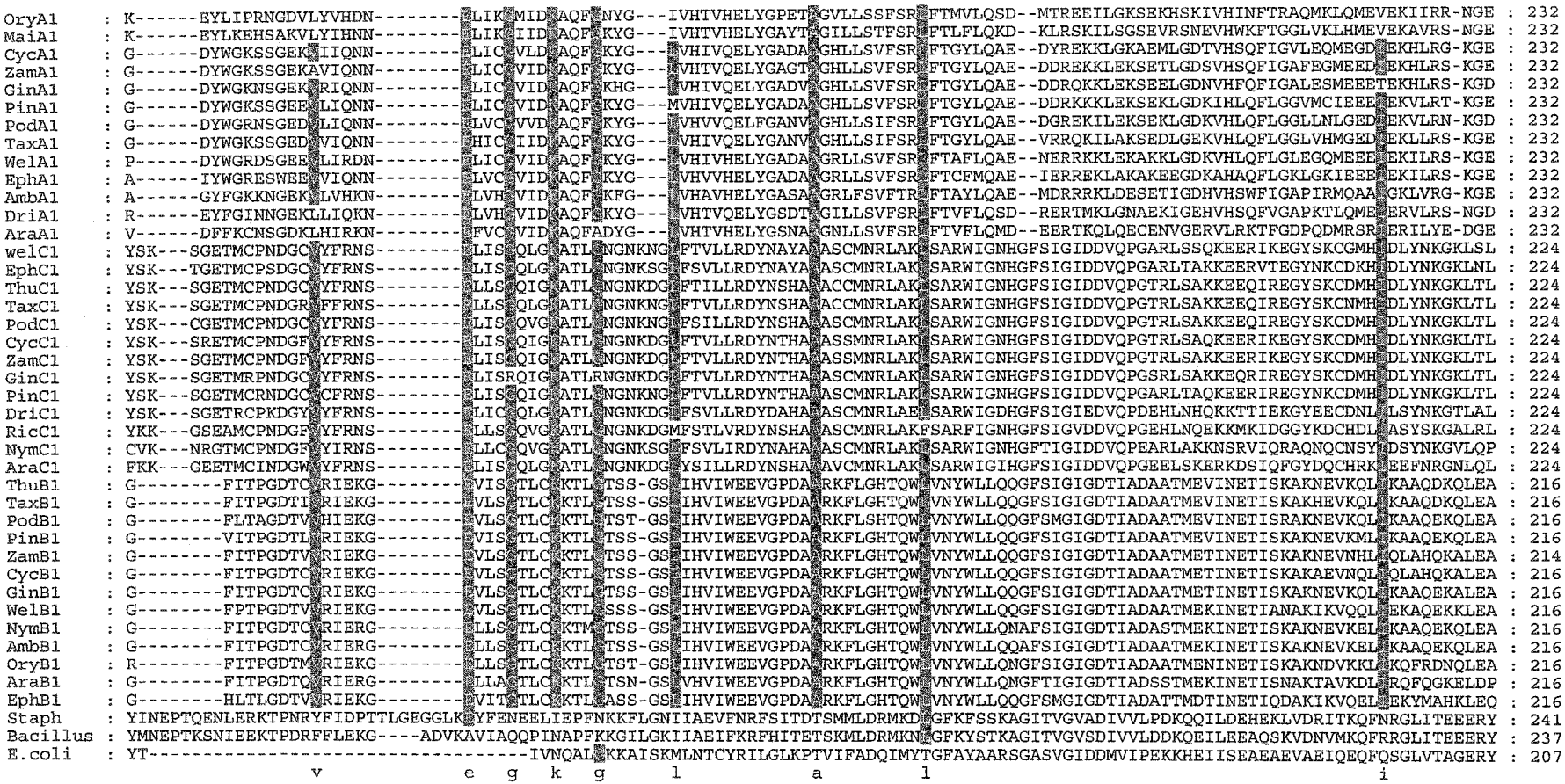
Appendix A4. Alignment of the translated amino acid sequences of *rpa1*, *rpb1* and *rpc1* genes and three bacterial β' genes. The alignment is shaded for three levels of conservation. Note that this alignment does not include the indels in each *rpa1*, *rpb1* and *rpc1* gene. See Table 1 for abbreviations.



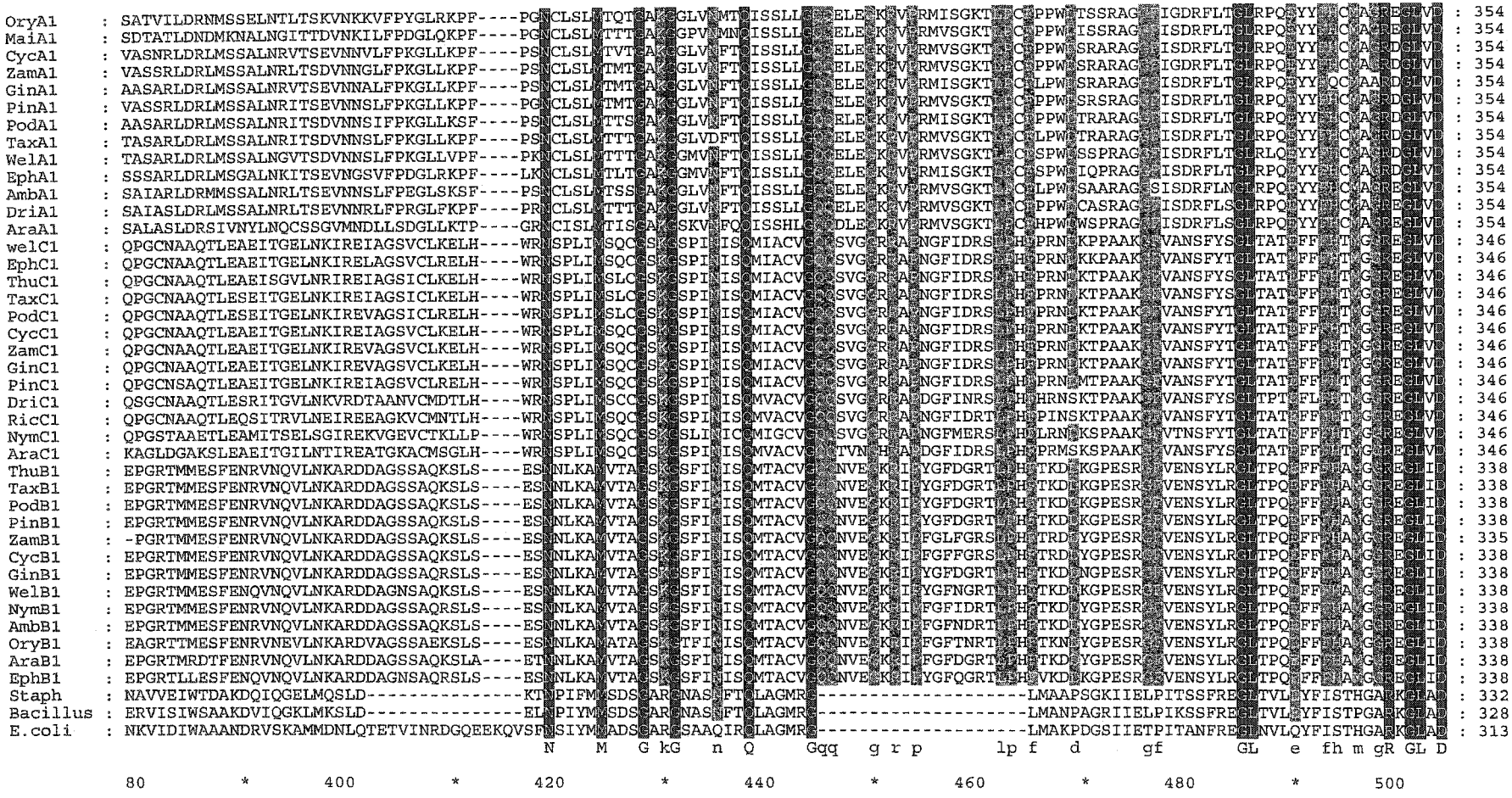
Appendix A4. Alignment of the translated amino acid sequences of *rpa1*, *rpb1* and *rpc1* genes and three bacterial β' genes. The alignment is shaded for three levels of conservation. Note that this alignment does not include the indels in each *rpa1*, *rpb1* and *rpc1* gene. See Table 1 for abbreviations.



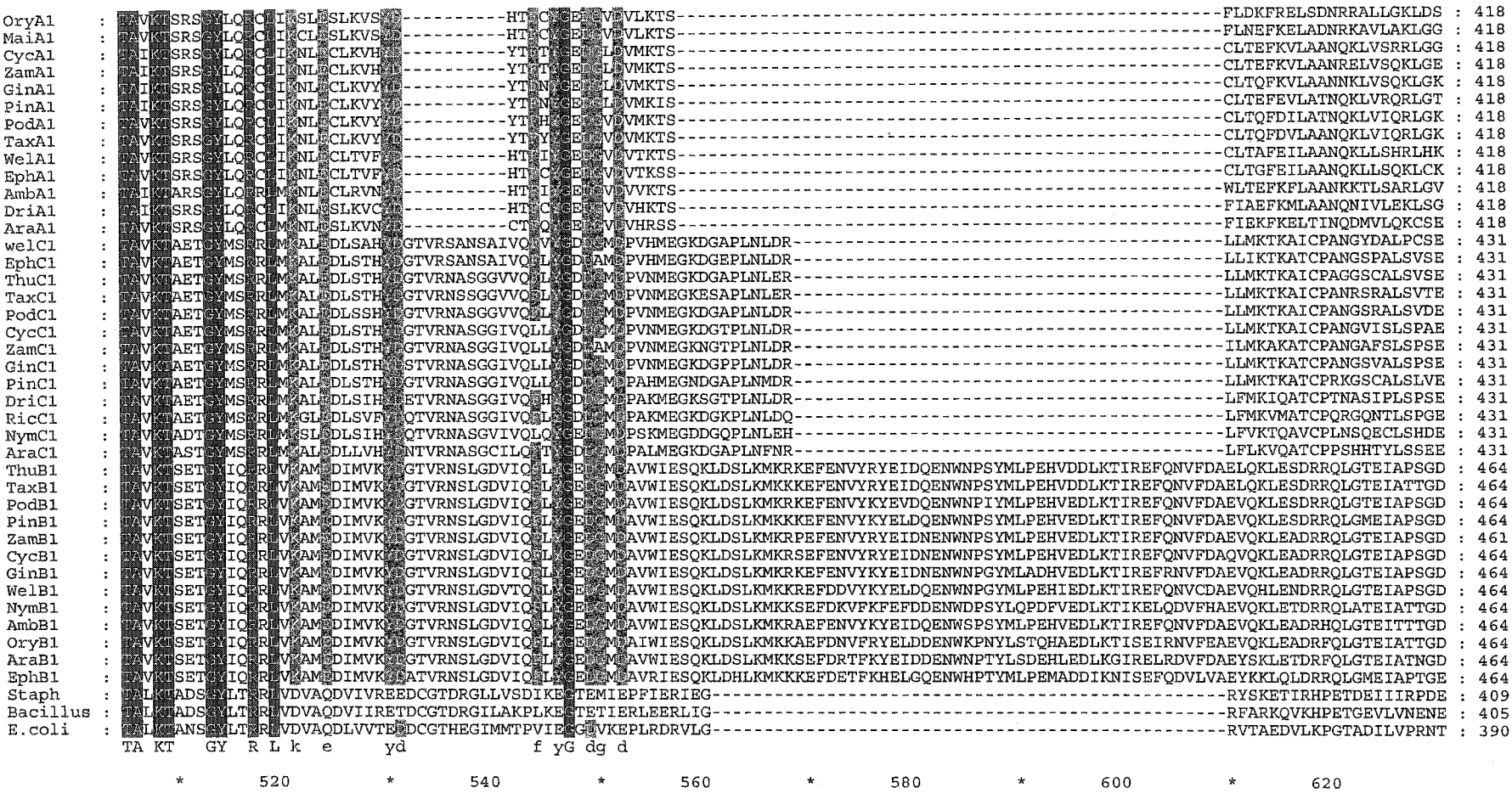
Appendix A4. *Continued*



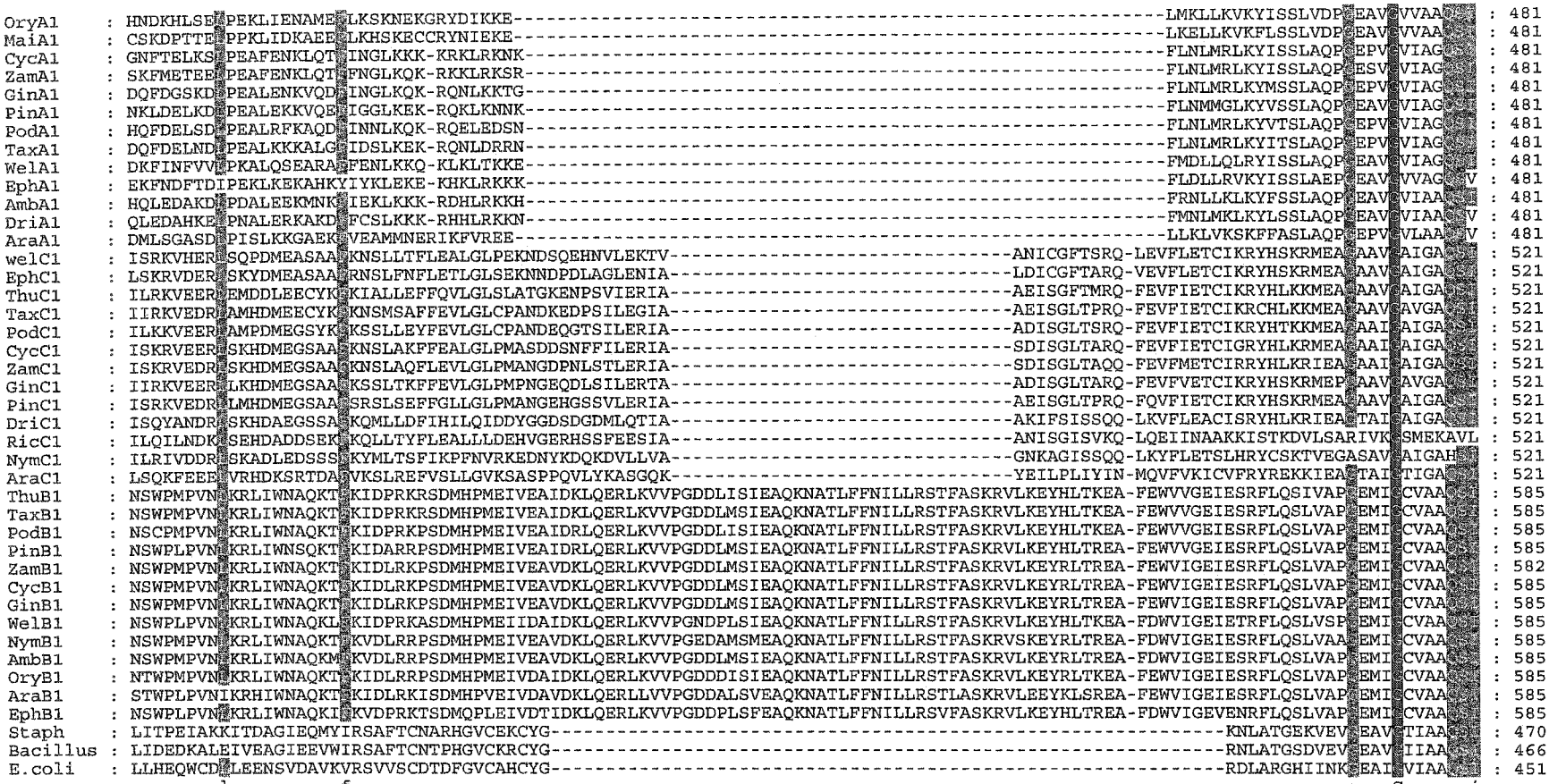
Appendix A4. *Continued*



Appendix A4. *Continued*



Appendix A4. *Continued*



Appendix B1. Non-synonymous (Ka) and synonymous (Ks) pairwise distances for *rpa1*. Ks and Ka were calculated using the method of Pamilio-Bianchi-Li (Pamilo and Bianchi 1993). Standard errors, calculated by bootstrapping 100 replicates, are shown in shaded columns. See Table 1 for abbreviations.

Ka for *rpa1*

	OryA1		MaiA1		CycA1		ZamA1		GinA1		PinA1		PodA1		TaxA1		WelA1		EphA1		AmbA1		DriA1	
OryA1																								
MaiA1	0.173	±0.014																						
CycA1	0.311	±0.030	0.342	±0.103																				
ZamA1	0.316	±0.089	0.344	±0.140	0.060	±0.008																		
GinA1	0.323	±0.068	0.340	±0.132	0.085	±0.011	0.08	±0.010																
PinA1	0.314	±0.158	0.338	±0.162	0.099	±0.011	0.102	±0.010	0.083	±0.010														
PodA1	0.331	±0.154	0.344	±0.087	0.138	±0.013	0.139	±0.014	0.113	±0.012	0.100	±0.011												
TaxA1	0.320	±0.126	0.330	±0.105	0.145	±0.015	0.139	±0.014	0.117	±0.013	0.117	±0.012	0.092	±0.012										
WelA1	0.332	±0.151	N/A		0.180	±0.015	0.194	±0.015	0.185	±0.015	0.164	±0.015	0.190	±0.014	0.201	±0.038								
EphA1	0.345	±0.175	0.363	±0.091	0.200	±0.018	0.220	±0.020	0.206	±0.019	0.196	±0.019	0.210	±0.018	0.216	±0.016	0.155	±0.014						
AmbA1	0.314	±0.023	0.343	±0.079	0.208	±0.017	0.205	±0.016	0.197	±0.017	0.202	±0.016	0.231	±0.016	0.247	±0.020	0.250	±0.019	0.273	±0.004				
DriA1	0.245	±0.020	0.273	±0.060	0.227	±0.030	0.228	±0.030	0.216	±0.035	0.206	±0.039	0.240	±0.029	0.255	±0.018	0.255	±0.054	0.293	±0.126	0.188	±0.015		
AraA1	0.410	±0.106	0.417	±0.172	0.366	±0.186	N/A		0.360	±0.183	0.359	±0.187	0.382	±0.145	0.375	±0.188	0.388	±0.059	0.394	±0.089	0.375	±0.155	0.319	±0.051

Ks for *rpa1*

	OryA1		MaiA1		CycA1		ZamA1		GinA1		PinA1		PodA1		TaxA1		WelA1		EphA1		AmbA1		DriA1	
OryA1																								
MaiA1	0.483	±0.048																						
CycA1	1.743	±0.453	1.880	±0.629																				
ZamA1	1.757	±0.587	1.961	±0.703	0.300	±0.035																		
GinA1	1.722	±0.363	1.972	±0.643	0.508	±0.057	0.559	±0.056																
PinA1	2.458	±1.146	2.177	±0.992	0.677	±0.066	0.646	±0.059	0.617	±0.062														
PodA1	2.185	±1.036	1.832	±0.508	0.753	±0.079	0.752	±0.066	0.664	±0.067	0.610	±0.063												
TaxA1	1.999	±0.872	1.924	±0.752	0.794	±0.072	0.802	±0.075	0.742	±0.063	0.727	±0.070	0.498	±0.046										
WelA1	2.410	±1.067	N/A	±1.150	1.243	±0.149	1.318	±0.166	1.168	±0.150	1.173	±0.121	1.356	±0.200	1.394	±0.314								
EphA1	2.692	±1.136	1.629	±0.428	1.219	±0.130	1.324	±0.174	1.325	±0.168	1.058	±0.116	1.155	±0.111	1.205	±0.183	0.982	±0.083						
AmbA1	1.464	±0.207	1.714	±0.407	1.202	±0.129	1.155	±0.117	1.052	±0.114	1.272	±0.133	1.309	±0.150	1.258	±0.137	1.570	±0.239	1.741	±0.514				
DriA1	1.372	±0.473	1.574	±0.431	1.370	±0.245	1.347	±0.135	1.322	±0.210	1.582	±0.237	1.551	±0.272	1.328	±0.225	1.753	±0.487	2.061	±0.890	1.216	±0.159		
AraA1	1.892	±0.645	1.963	±0.854	2.297	±1.119	N/A	±1.095	2.388	±1.171	2.188	±0.974	1.739	±0.601	2.484	±1.177	1.717	±0.422	1.732	±0.435	2.099	±0.884	1.574	±0.311

Appendix B2. Non-synonymous (Ka) and synonymous (Ks) pairwise distances for *rpb1*. Ks and Ka were calculated using the method of Pamilio-Bianchi-Li (Pamilo and Bianchi 1993). Standard errors, calculated by bootstrapping 100 replicates, are shown in shaded columns. See Table 1 for abbreviations.

Ka for *rpb1*

	ThuB1		TaxB1		PodB1		PinB1		ZamB1		CycB1		GinB1		NymB1		AmbB1		OryA1		ArabA1		WelA1	
ThuB1																								
TaxB1	0.008	±0.002																						
PodB1	0.017	±0.004	0.020	±0.004																				
PinB1	0.025	±0.005	0.025	±0.005	0.024	±0.004																		
ZamB1	0.035	±0.006	0.036	±0.006	0.038	±0.005	0.043	±0.006																
CycB1	0.033	±0.005	0.036	±0.005	0.040	±0.006	0.044	±0.006	0.007	±0.003														
GinB1	0.027	±0.005	0.030	±0.006	0.032	±0.006	0.035	±0.006	0.021	±0.004	0.021	±0.004												
NymB1	0.064	±0.011	0.064	±0.008	0.063	±0.010	0.065	±0.011	0.055	±0.006	0.056	±0.008	0.049	±0.006										
AmbB1	0.043	±0.008	0.044	±0.007	0.043	±0.007	0.043	±0.007	0.033	±0.005	0.034	±0.005	0.029	±0.005	0.037	±0.005								
OryB1	0.073	±0.026	0.071	±0.013	0.070	±0.019	0.075	±0.029	0.068	±0.011	0.070	±0.007	0.069	±0.034	0.071	±0.017	0.065	±0.007						
ArabB1	N/A		N/A		N/A		N/A		0.091	±0.035	0.093	±0.031	0.091	±0.034	0.089	±0.036	0.087	±0.009	0.093	±0.029				
WelB1	0.053	±0.007	0.053	±0.007	0.049	±0.007	0.055	±0.007	0.063	±0.008	0.062	±0.006	0.055	±0.007	0.084	±0.009	0.064	±0.008	0.091	±0.022	0.110	±0.035		
EpiB1	0.094	±0.009	0.094	±0.009	0.093	±0.012	0.097	±0.010	0.109	±0.010	0.111	±0.013	0.101	±0.010	0.114	±0.011	0.107	±0.046	0.123	±0.024	0.129	±0.050	0.084	±0.008

Ks for *rpb1*

	ThuB1		TaxB1		PodB1		PinB1		ZamB1		CycB1		GinB1		NymB1		AmbB1		OryA1		ArabA1		WelA1	
ThuB1																								
TaxB1	0.298	±0.014																						
PodB1	0.559	±0.053	0.546	±0.048																				
PinB1	0.652	±0.057	0.565	±0.049	0.582	±0.052																		
ZamB1	0.850	±0.083	0.774	±0.083	0.803	±0.071	0.687	±0.065																
CycB1	0.794	±0.072	0.758	±0.062	0.791	±0.073	0.66	±0.068	0.276	±0.031														
GinB1	0.772	±0.081	0.727	±0.078	0.806	±0.070	0.625	±0.063	0.476	±0.047	0.470	±0.041												
NymB1	1.693	±0.433	1.499	±0.291	1.760	±0.530	1.506	±0.269	1.272	±0.123	1.375	±0.167	1.523	±0.218										
AmbB1	1.560	±0.357	1.372	±0.216	1.585	±0.362	1.436	±0.232	1.170	±0.120	1.168	±0.113	1.152	±0.117	1.240	±0.117								
OryB1	1.896	±0.880	1.885	±0.750	1.794	±0.861	1.951	±0.893	1.747	±0.494	1.555	±0.280	2.497	±1.031	1.710	±0.546	1.219	±0.152						
ArabB1	N/A		N/A		N/A		N/A		1.877	±0.724	1.847	±0.581	1.931	±0.692	2.076	±0.872	1.564	±0.244	1.903	±0.643				
WelB1	1.344	±0.146	1.201	±0.117	1.265	±0.143	0.999	±0.102	1.072	±0.105	0.986	±0.095	1.066	±0.089	1.525	±0.161	1.192	±0.122	1.786	±0.537	1.966	±0.818		
EpiB1	1.229	±0.148	1.158	±0.130	1.397	±0.180	1.161	±0.129	1.295	±0.163	1.347	±0.213	1.430	±0.241	1.546	±0.208	2.049	±0.917	1.854	±0.553	2.011	±0.889	1.385	±0.204

Appendix B3. Non-synonymous (Ka) and synonymous (Ks) pairwise distances for *rpc1*. Ks and Ka were calculated using the method of Pamilio-Bianchi-Li (Pamilo and Bianchi 1993). Standard errors, calculated by bootstrapping 100 replicates, are shown in shaded columns. See Table 1 for abbreviations.

Ka for *rpc1*

	WelC1		EphC1		ThuC1		TaxC1		PodC1		PinC1		CycC1		ZamC1		GinC1		AraC1		NymC1		DriC1	
WelC1																								
EphC1	0.075	±0.009																						
ThuC1	0.102	±0.011	0.111	±0.011																				
TaxC1	0.097	±0.011	0.109	±0.010	0.042	±0.007																		
PodC1	0.093	±0.011	0.103	±0.009	0.052	±0.008	0.043	±0.006																
PinC1	0.091	±0.011	0.099	±0.010	0.081	±0.009	0.078	±0.009	0.072	±0.008														
CycC1	0.087	±0.011	0.100	±0.011	0.087	±0.010	0.083	±0.010	0.079	±0.009	0.076	±0.009												
ZamC1	0.095	±0.015	0.096	±0.010	0.086	±0.011	0.087	±0.010	0.082	±0.010	0.074	±0.008	0.027	±0.005										
GinC1	0.080	±0.010	0.099	±0.010	0.076	±0.009	0.070	±0.010	0.067	±0.010	0.060	±0.007	0.049	±0.006	0.045	±0.006								
AraC1	N/A		0.267	±0.114	0.258	±0.120	0.266	±0.099	0.246	±0.119	N/A		0.244	±0.073	0.245	±0.101	0.251	±0.122						
NymC1	0.226	±0.019	0.229	±0.037	0.241	±0.032	0.231	±0.031	0.234	±0.029	0.232	±0.029	0.223	±0.037	0.231	±0.029	0.220	±0.029	N/A					
DriC1	0.193	±0.041	0.202	±0.036	0.202	±0.033	0.205	±0.018	0.200	±0.026	0.191	±0.025	0.177	±0.016	0.178	±0.016	0.188	±0.017	0.263	±0.021	0.240	±0.020		
DryC1	N/A		0.270	±0.113	0.256	±0.105	0.257	±0.122	0.245	±0.087	0.257	±0.055	0.243	±0.062	0.246	±0.071	0.251	±0.096	0.303	±0.138	0.281	±0.138	0.211	±0.057

Ks for *rpc1*

	WelC1		EphC1		ThuC1		TaxC1		PodC1		PinC1		CycC1		ZamC1		GinC1		AraC1		NymC1		DriC1	
WelC1																								
EphC1	1.080	±0.108																						
ThuC1	1.229	±0.138	1.193	±0.115																				
TaxC1	1.329	±0.140	1.198	±0.112	0.320	±0.039																		
PodC1	1.454	±0.199	1.115	±0.100	0.547	±0.057	0.410	±0.044																
PinC1	1.151	±0.126	1.156	±0.126	0.648	±0.066	0.512	±0.052	0.596	±0.059														
CycC1	1.388	±0.150	1.210	±0.112	0.752	±0.081	0.737	±0.075	0.736	±0.069	0.659	±0.067												
ZamC1	1.281	±0.162	1.357	±0.163	0.710	±0.073	0.755	±0.077	0.800	±0.078	0.727	±0.069	0.302	±0.035										
GinC1	1.443	±0.152	1.251	±0.123	0.663	±0.052	0.671	±0.056	0.624	±0.056	0.591	±0.054	0.471	±0.050	0.537	±0.049								
AraC1	N/A		2.167	±0.870	2.191	±0.975	2.066	±0.816	2.553	±1.195	N/A		1.955	±0.738	2.034	±0.861	2.529	±1.196						
NymC1	1.388	±0.166	1.769	±0.679	1.623	±0.306	1.511	±0.244	1.594	±0.213	1.401	±0.191	1.577	±0.400	1.485	±0.206	1.431	±0.184	N/A					
DriC1	1.721	±0.381	1.569	±0.101	1.599	±0.335	1.416	±0.183	1.534	±0.345	1.448	±0.171	1.302	±0.150	1.292	±0.141	1.237	±0.170	1.499	±0.250	0.961	±0.090		
DryC1	N/A		2.021	±0.803	2.073	±0.905	2.432	±1.168	1.951	±0.712	1.767	±0.420	1.666	±0.360	1.859	±0.559	1.906	±0.748	2.157	±1.004	2.086	±0.477	1.509	±0.306

Appendix B4. Pairwise amino acid distances (Kaa) for *rpa1*, *rpb1* and *rpc1*. Amino acid distances were calculated according to a Poisson correction. Standard errors, calculated by bootstrapping 100 replicates, are shown in shaded columns. See Table 1 for abbreviations.

Kaa for *rpa1*

	OryA1		*ZeaA1*		CycA1		ZamA1		GinA1		PinA1		PodA1		TaxA1		WelA1		EphA1		AmbA1		DriA1	
OryA1																								
ZeaA1	0.290	±0.025																						
CycA1	0.497	±0.035	0.514	±0.038																				
ZamA1	0.480	±0.035	0.490	±0.037	0.107	±0.015																		
GinA1	0.525	±0.037	0.511	±0.038	0.151	±0.018	0.151	±0.019																
PinA1	0.497	±0.035	0.490	±0.037	0.178	±0.020	0.190	±0.021	0.163	±0.018														
PodA1	0.487	±0.034	0.490	±0.037	0.221	±0.022	0.218	±0.022	0.195	±0.020	0.193	±0.020												
TaxA1	0.507	±0.035	0.467	±0.037	0.226	±0.022	0.221	±0.024	0.203	±0.021	0.215	±0.022	0.161	±0.018										
WelA1	0.494	±0.035	0.494	±0.035	0.293	±0.026	0.301	±0.025	0.306	±0.026	0.282	±0.025	0.315	±0.027	0.315	±0.026								
EphA1	0.494	±0.035	0.494	±0.036	0.326	±0.027	0.344	±0.028	0.338	±0.027	0.323	±0.027	0.344	±0.029	0.355	±0.028	0.244	±0.023						
AmbA1	0.484	±0.035	0.518	±0.036	0.329	±0.028	0.341	±0.029	0.329	±0.028	0.355	±0.029	0.373	±0.029	0.400	±0.031	0.376	±0.031	0.410	±0.032				
DriA1	0.391	±0.030	0.406	±0.030	0.361	±0.029	0.364	±0.030	0.379	±0.030	0.364	±0.029	0.391	±0.030	0.413	±0.031	0.391	±0.030	0.432	±0.032	0.326	±0.027		
ArzA1	0.586	±0.035	0.564	±0.036	0.564	±0.036	0.549	±0.036	0.567	±0.039	0.564	±0.040	0.560	±0.038	0.549	±0.039	0.586	±0.039	0.578	±0.039	0.553	±0.037	0.461	±0.034

Kaa for *rpb1*

	ThuB1		TaxB1		PodB1		PinB1		ZamB1		CycB1		GinB1		NymB1		AmbB1		OryA1		ArabA1		WelA1	
ThuB1																								
TaxB1	0.015	±0.005																						
PodB1	0.034	±0.007	0.041	±0.008																				
PinB1	0.039	±0.008	0.041	±0.008	0.046	±0.009																		
ZamB1	0.051	±0.009	0.055	±0.009	0.064	±0.010	0.064	±0.010																
CycB1	0.051	±0.009	0.055	±0.009	0.067	±0.010	0.067	±0.010	0.012	±0.004														
GinB1	0.043	±0.008	0.048	±0.009	0.051	±0.009	0.053	±0.010	0.036	±0.007	0.036	±0.007												
NymB1	0.102	±0.013	0.096	±0.013	0.107	±0.013	0.100	±0.014	0.096	±0.013	0.096	±0.013	0.085	±0.012										
AmbB1	0.060	±0.010	0.058	±0.010	0.067	±0.011	0.062	±0.011	0.055	±0.009	0.057	±0.009	0.051	±0.009	0.060	±0.010								
OryB1	0.115	±0.014	0.109	±0.014	0.120	±0.015	0.120	±0.015	0.115	±0.014	0.119	±0.014	0.113	±0.014	0.119	±0.014	0.106	±0.013						
AraB1	0.147	±0.016	0.141	±0.016	0.153	±0.016	0.155	±0.016	0.143	±0.016	0.143	±0.016	0.138	±0.015	0.136	±0.015	0.138	±0.016	0.147	±0.016				
WelB1	0.093	±0.012	0.093	±0.012	0.091	±0.012	0.094	±0.012	0.106	±0.013	0.106	±0.013	0.096	±0.012	0.140	±0.015	0.106	±0.013	0.153	±0.016	0.171	±0.017		
EphB1	0.153	±0.018	0.153	±0.018	0.155	±0.018	0.151	±0.018	0.169	±0.018	0.173	±0.018	0.165	±0.018	0.179	±0.017	0.167	±0.018	0.197	±0.019	0.201	±0.018	0.138	±0.014

Appendix B4. ***Continued.***

Kaa for *rpc1*

	WelC1	EphC1	ThuC1	TaxC1	PodC1	PinC1	CycC1	ZamC1	GinC1	AraC1	NymC1	DriC1
welC1												
EphC1	0.130 ±0.015											
ThuC1	0.165 ±0.018	0.194 ±0.020										
TaxC1	0.161 ±0.017	0.197 ±0.019	0.084 ±0.013									
PodC1	0.148 ±0.017	0.172 ±0.017	0.099 ±0.013	0.088 ±0.013								
PinC1	0.145 ±0.017	0.170 ±0.018	0.145 ±0.017	0.145 ±0.016	0.132 ±0.016							
CycC1	0.143 ±0.017	0.170 ±0.017	0.148 ±0.017	0.145 ±0.016	0.126 ±0.015	0.128 ±0.015						
ZamC1	0.165 ±0.016	0.165 ±0.017	0.159 ±0.018	0.161 ±0.017	0.141 ±0.016	0.137 ±0.016	0.058 ±0.010					
GinC1	0.145 ±0.017	0.167 ±0.017	0.141 ±0.016	0.135 ±0.016	0.113 ±0.015	0.111 ±0.013	0.095 ±0.013	0.092 ±0.013				
AraC1	0.369 ±0.029	0.393 ±0.028	0.363 ±0.029	0.377 ±0.028	0.358 ±0.028	0.360 ±0.028	0.360 ±0.028	0.366 ±0.028	0.371 ±0.028			
NymC1	0.355 ±0.026	0.369 ±0.028	0.379 ±0.030	0.379 ±0.029	0.369 ±0.029	0.371 ±0.029	0.355 ±0.027	0.366 ±0.028	0.350 ±0.028	0.454 ±0.032		
DriC1	0.300 ±0.020	0.313 ±0.020	0.318 ±0.021	0.318 ±0.020	0.303 ±0.020	0.300 ±0.020	0.278 ±0.023	0.275 ±0.023	0.290 ±0.024	0.388 ±0.030	0.379 ±0.029	
OryC1	0.369 ±0.028	0.391 ±0.028	0.360 ±0.029	0.369 ±0.028	0.344 ±0.027	0.350 ±0.026	0.350 ±0.027	0.352 ±0.027	0.350 ±0.026	0.439 ±0.032	0.430 ±0.031	0.350 ±0.028

Selecting genes for phylogenetic analysis

The issue of selecting genes (sequence information) for phylogenetic analysis is important but not very well studied. Traditionally, workers have used the sequences of genes that were easier to obtain such as rRNA genes (which are present in numerous copies in most genomes). However, a single gene cannot usually provide phylogenetic resolution for different divergence levels. In addition, a single gene phylogeny might be misleading because of biases and peculiarities that can influence the phylogenetic signal in a certain gene. For example, it has been shown that organism specific molecular evolution, such as GC bias can influence phylogenies at DNA and amino acid levels (e.g., Foster and Hickey 1999). The issue of selecting genes is also coupled with the limitations in the methods of analysis. For example, the popular method of maximum parsimony can be misleading when the sequences evolve rapidly and undergo multiple substitutions.

A reasonable strategy for addressing phylogenies has been to combine sequences of different genes and to make large data sets with phylogenetic signal for different divergence levels. This approach has been helpful in resolving several phylogenetic issues in recent years. However, there are still no objective criteria for selecting and combining genes for phylogenetic analysis. Normalized splits (NS), is an attempt in using spectral analysis for selecting and combining genes and gene partitions for addressing a certain phylogenetic problem. This method is interesting because spectral analysis does not make several assumptions about the data and is done independently of any phylogenetic trees. Our results with this method show its effectiveness for finding genes with noisy data partitions. A method like NS can give us an initial assessment of the noise in the genes and their partitions. We can later decide on choosing data partitions

that are less noisy or to use a method of analysis that better fits the sequences with high noise (e.g., maximum likelihood with a probabilistic model). We have tested the NS method in two phylogenetic analyses of plants with successful results (Chapters 3 and 4). Testing this approach in other phylogenetic systems, especially those with known branching orders, can help in further evaluating this approach.

Criticizing NS

A potential problem with NS is caused when there are a large number of invariant sites in a sequence data set. For example for 4 taxa a data set containing 5 variable sites (producing 5 splits) and 495 invariant sites gives an NS of 0.01 whereas a sequence data with 2 variable sites (producing 2 splits) and 18 invariable sites gives an NS of 0.1. Although the former sequence data is noisier than the later, but its invariable sites can lead a smaller NS (lower noise). To address this problem we used two other measures: calculating NS by dividing the number of splits by the number of variable sites (NS-V), as the source of splits, and using the number of splits from each data set without normalizing them. Figure 1 compares the result we obtained using these new measures using a data set of 9 genes (and their partitions) used in Chapter 4 for seed plant phylogeny. NS-V did not seem to vary greatly between genes and partitions. It is unlikely that this measure can be useful for selecting genes and partitions for phylogenetic studies. On the other hand, the total number of splits, greatly varies between different data sets and data partitions. For example the third codon positions of nuclear RNA polymerase genes show a much higher number of splits and possibly a greater level of noise

comparing to first and second codon positions. This is in agreement with a higher rate of substitution in third base of codons (see Chapter 4).

Building a data set for addressing seed plant phylogeny

Most of the genes that have been used in addressing plant phylogeny are from organellar genomes or ribosomal genes. RNA polymerases were selected as a new line of evidence and to address the lack of an extensive data set from nuclear protein-coding genes. Other studies had shown that RNA polymerases are excellent phylogenetic markers (i.e. Nickerson and Drouin in press; Stiller and Hall 1997; Hirt et al. 1999; Stiller, Riley, and Hall 2001). We selected the largest subunit of three nuclear RNA polymerases I, II and III (*rpa1, rpb1* and *rpc1*) because these genes evolve with different rates, which allows covering a wide divergence range. These three genes are also technically easy to PCR amplify, clone and sequence. The focus of our sampling was to obtain *rpa1, rpb1* and *rpc1* genes from all seed plant groups to address seed plant phylogeny, in general, and to assess the recently proposed Gnepines hypothesis, which places Gnetales within conifers and as a sister group to Pinaceae. Therefore, we sequenced and cloned these genes in representatives of all major groups of seed plants (angiosperms, and four gymnosperms: conifers, cycads, Ginkgo and Gnetales) as well as 4 major groups of conifers (*Taxus, Podocarpus, Thuja* and *Pinus*) and two groups of Gnetales (*Welwitschia* and *Ephedra*). To compare and combine with RNA polymerase data set, we also gathered available sequence information from chloroplast, mitochondria and 18S rRNA genes.

While the focus of this work was major seed plant groups, our study offers insight in the phylogenetic utility of *rpa1*, *rpb1* and *rpc1*. Our evolutionary rate analysis confirms that *rpb1* is the slowest evolving of these three genes. Therefore, its sequence information should be useful in resolving deep branches of plant phylogeny such as the relationships of liverworths, ferns, gymnosperms and angiosperms. We found that *rpc1* and *rpa1* evolve faster than *rpb1*, and their sequences should be more useful than *rpb1* in resolving relationships of closely related groups (such as different angiosperms groups which diverged less than 100MYA).

Seed plant phylogeny and the position of Gnetales resolved

The major phylogenetic question that was addressed in this thesis was the position of Gnetales. This issue has been debated for several decades and amid several recent molecular studies, it is still controversial. Our phylogenetic analyses provided a well-supported phylogeny for seed plants. In this tree, gymnosperms and angiosperms are monophyletic, the Gnepines hypothesis is supported (Gnetales are nested within conifers, as a sister group to Pinaceae), cycads are the first branch of gymnosperms, and Ginkgo is grouped with a branch leading to conifers+Gnetales. Using NS analysis and evolutionary rates, we found that the third codon positions of RNA polymerase genes and chloroplast genes have encountered multiple substitutions at the divergence level of seed plant phylogeny. Therefore, these data partitions are not appropriate for resolving seed plant phylogeny especially in the maximum parsimony criterion (see above). Comparing this observation with recent studies in plant phylogeny make us believe that the choice of genes (or partitions of genes) in combination with the method of analysis are the major

causes of discrepancies in recent phylogenetic studies of seed plants. As was originally proposed by Felsenstein (1978), we firmly believe that the use of maximum parsimony should be cautioned when data partitions have encountered multiple substitutions.

Molecular evolution of RNA polymerases

We addressed molecular evolution of the largest subunit of RNA polymerases using the sequence data set of *rpa1, rpb1* and *rpc1* in seed plants. A phylogeny of these genes revealed that they share a common ancestor and that *rpc1* and *rpb1* are closer. We showed that these genes evolve with different rates at non-synonymous sites and amino acid level. *rpa1* evolves faster followed by *rpc1* and *rpb1*. The evolutionary rates are slower in the regions that have important functional properties. We also studied the level of functional constraint in each amino acid position of these molecules using a maximum likelihood analysis of the ratio of non-synonymous to synonymous substitutions (Yang and Bielawski 2000). We compared the pattern of evolutionary functional constraint with functional properties of the same region of RNAP (known from the crystallographic structure). Interestingly, our evolutionary map corresponds to the structural map. Regions with important functions are under higher negative selection. However, the level of negative selection is lower in *rpa1* and *rpc1* than in *rpb1*. Our data set included a functionally important part of these genes (region D-G; see Chapter 1, Figure 1) in seed plants. It will be interesting to study the molecular evolution of these genes using the sequence of complete genes and in other groups of organisms.

Is the concerted evolution a driving force in the evolution of *rpa1, rpb1* and *rpc1*?

Concerted evolution is a common phenomenon in rRNA and tRNA genes. It is caused by recombination and it leads to homogenization of rRNA and tRNA genes in the genome of a species. This makes the promoter signals of the rRNA and tRNA genes evolve faster. However, concerted evolution does not accelerate the evolution of promoters of protein-coding genes because their promoter signals are found upstream of different genes. Dover and Flavell (1984) suggested that RNAP I has to adapt to the accelerated rate of evolution of rRNA promoter brought about by concerted evolution. If this idea is correct, the proteins involved in transcribing rRNA and tRNA genes, are therefore expected to evolve faster than those involved in transcribing protein coding genes. It has been shown that the transcription factors of RNAP I are species-specific (Heix and Grummt 1995). Our study showed that the evolutionary rates of *rpa1, rpb1* and *rpc1* are positively related with the amount of concerted evolution in the genes they transcribe. This suggests that *rpa1* and *rpc1* evolve (change) more rapidly than *rpb1* to adapt to the concerted changes in the promoter regions of the genes they transcribe.

References

Dover, G., and R. B. Flavell. 1984. Molecular coevolution: DNA divergence and maintenance of function. Cell **38**:622-623.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**:401-410.

Foster, P. G., and D. A. Hickey. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. **48**:284-290.

Heix, J., and I. Grummt. 1995. Species specificity of transcription by RNA polymerase I. Curr. Opin. Genet. Dev. **5**:652-656.

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

Nickerson, J. and G. Drouin. in press. The sequence of the largest subunit of RNA polymerase II is a useful marker for inferring seed plant phylogeny. Mol. Phylogenet. Evol.

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

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

Yang, Z., and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. **15**:496-503.

Figure 1. Comparison of normalized splits (NS, panel A), normalized splits using variable sites (NS-V, panel B) and total number of splits (panel C) from 9 genes and their partitions. NS, NS-V and number of splits were calculated using the direct encoding method in the Spectronet program.

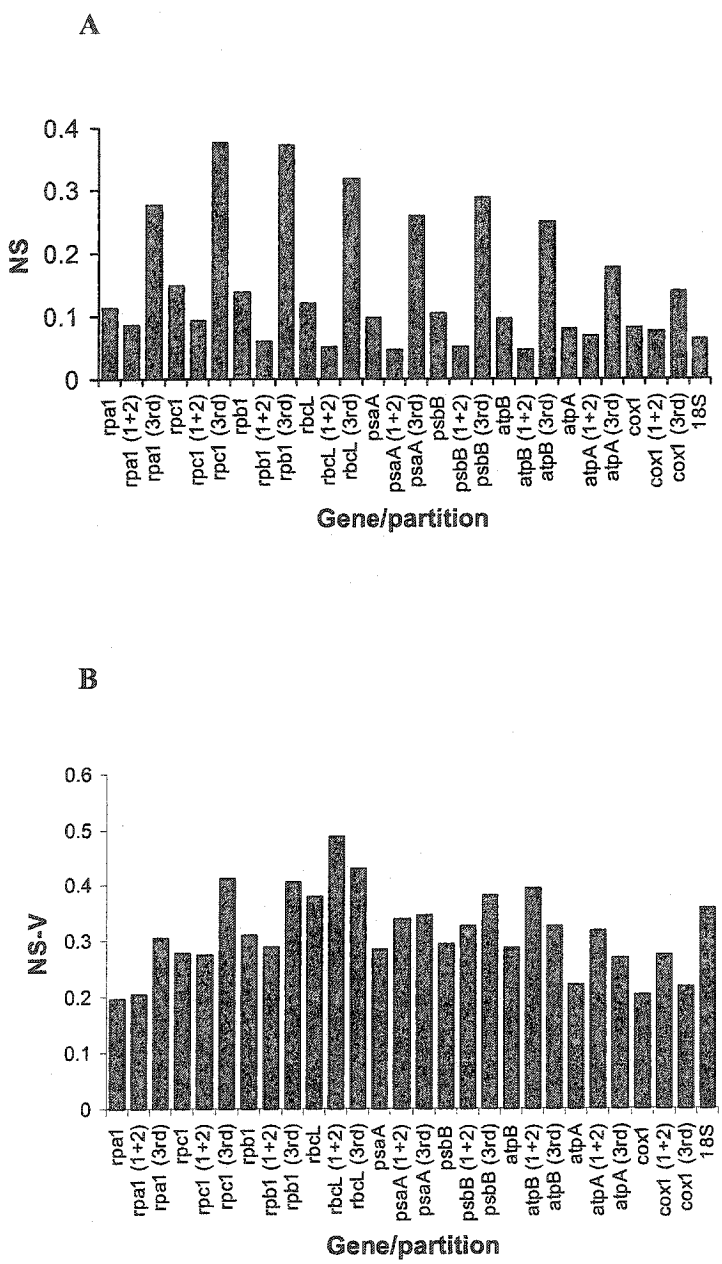


Figure 1. *Continued*

C