LARGE TREES, SUPERTREES AND THE GRASS PHYLOGENY

Thesis submitted to the University of Dublin, Trinity College
for the
Degree of Doctor of Philosophy (Ph.D.)

by

Nicolas Salamin

Department of Botany
University of Dublin, Trinity College

2002

Research conducted under the supervision of
Dr. Trevor R. Hodkinson
Department of Botany, University of Dublin, Trinity College

Dr. Vincent Savolainen
Jodrell Laboratory, Molecular Systematics Section,
Royal Botanic Gardens, Kew, London

DECLARATION

I thereby certify that this thesis has not been submitted as an exercise for a degree at any other University. This thesis contains research based on my own work, except where otherwise stated.

I grant full permission to the Library of Trinity College to lend or copy this thesis upon request.

SIGNED: ________________________

ACKNOWLEDGMENTS

I wish to thank Trevor Hodkinson and Vincent Savolainen for all the encouragement they gave me during the last three years. They provided very useful advice on scientific papers, presentation lectures and all aspects of the supervision of this thesis.

It has been a great experience to work in Ireland, and I am especially grateful to Trevor for the warm welcome and all the help he gave me, at work or outside work, since the beginning of this Ph.D. in the Botany Department. I will always remember his patience and kindness to me at this time.

I am also grateful to Vincent for his help and warm welcome during the different periods of time I stayed in London, but especially for all he did for me since my B.Sc. at the University of Lausanne.

I wish also to thank Prof. Mark Chase for his comments and advice on different manuscripts taken from this thesis.

Thanks to Jonathan Davis for being an interested beta tester for my programs, and the members of the Molecular Systematic Section from the Jodrell Laboratory for their welcome.

Thanks finally to all members of the Botany Department from Trinity College. I will not forget Ireland!

Finally, I could not imagine having spent the last three years in Dublin without Karine.

This project was funded by the Irish Higher Education Authority, the Roche Research Foundation, the Agassiz Foundation and a University of Lausanne Research Grant.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

Phylogenetic inference

BR = Baum (1992), Ragan (1992) coding scheme for MRP supertree reconstruction
D1 = agreement subtree metric
GTR = general time reversible model of substitution
HKY85 = Hasegawa, Kishino and Yano (1985) model of substitution
K2P = Kimura (1980) two-parameters model of substitution
KH test = Shimodaira and Hasegawa (1999) test
Log-Det = log-determinant or paralinear distance
MCMC = Markov chain Monte Carlo
ME = minimum evolution
ML = maximum likelihood
MP = maximum parsimony
MRP = matrix representation using parsimony
NJ = neighbour-joining
NNI = nearest-neighbour interchange
NPRS = non parametric rate smoothing
PL = penalized likelihood
PM = partition metric or symmetric difference
PU = Purvis (1996) coding scheme for MRP supertree reconstruction
SH test = Kishino and Hasegawa (1989) test
SPR = subtree pruning and regrafting
TBR = tree bisection and reconnection
ti/tv ratio = ratio of transitions over transversions in a DNA sequence
TrN = Tamura and Nei (1993) model of substitution

DNA sequencing

1, 2, or 3 = when associated with a coding gene represent the first, second and third codon positions respectively. Similarly, 1+2 or 1+2+3 represent the first and second or the three codon positions.
18S rDNA = nuclear ribosomal gene coding for the 18S subunit
5.8S rDNA = nuclear ribosomal gene coding for the 5.8S subunit

atpB = ATP synthase beta subunit

bp = base pairs or number of nucleotides

gbss = granule bound starch synthase

ITS (1 or 2) = internal transcribed spacer (1 or 2) of 5.8S rDNA

ndhF = NADH dehydrogenase subunit F

phyB = phytochrome B

rbcL = ribulose-1,5-bisphosphate carboxylase large subunit

trnLF = partial Leucine transfer RNA (*trnL*) exons + *trnL* intron + *trnL-trnF* intergenic spacer + partial Phenylalanine transfer RNA (*trnF*) exon

Grass systematics

APG = angiosperm phylogeny group

BEP or BE-P = clade containing Bambusoideae, Ehrhartoideae and Pooideae. BE-P is used in chapter 5 to emphasize the position of the Pooideae.

GPWG = grass phylogeny working group

PACCAD = clade containing Panicoideae, Arundinoideae, Centothecoideae, Chloridoideae, Aristidoideae and Danthonioideae

PACCAD-P = clade containing PACCAD clade and Pooideae

s.l. = *sensu lato*

s.s. = *sensu stricto*

ABSTRACT

During the last decade, the advances of molecular techniques have profoundly changed the way scientists build and use phylogenetic trees. Vast fields of research as different as ecology, evolution of development, genomics, and systematics have been influenced by the growth of phylogenetics, and the possibilities offered by new techniques of tree reconstruction are likely to further anchor the discipline as a core component of evolutionary biology. Despite this, phylogenetic inference remains a particularly difficult task because no polynomial-time algorithm is available to reconstruct optimal trees based on a given data set and the problem is getting more difficult as the number of taxa handled in reconstructions increases. The last decade has witnessed the development of powerful computer architectures and software that have alleviated this burden. However, the reconstruction of comprehensive phylogenetic trees still has to rely on heuristic searches and sound statistical methods are often prohibited for large data sets due to associated computational difficulties.

In this thesis, I explored the problem of reconstructing large phylogenetic trees. One aspect was to investigate how well current methods of tree reconstruction performed when faced with matrices containing hundreds or thousands of taxa. In chapter 2 of this thesis, computer simulations based on four large angiosperm trees were performed to assess the success of maximum parsimony and neighbour-joining to infer trees. The results indicated that the size of the matrix was not a problem in itself, and that the distribution of changes along the tree could be a more important factor. For instance, when conditions were favourable, more than 80% of the nodes from a tree containing 13,000 taxa could be correctly inferred with simulated data sets of 10,000 bp. With real data sets, it is however impossible to know how far the trees obtained are from the ‘true’ underlying evolutionary hypothesis. Resampling techniques, such as bootstrap or jackknife, have been developed to estimate how much confidence one can put on a particular node of a phylogenetic tree. With large numbers of taxa, these procedures become computationally intensive, especially if thorough heuristic searches are used. It is therefore important to understand the effects of different heuristic strategies on the support obtained for large phylogenetic trees, and whether faster tree search options could be used to reduce the time of the analyses without biasing the support obtained. In chapter 3, the level of support obtained by bootstrapping and jackknifing a 357 taxa molecular matrix for the angiosperms using four different heuristic search options were compared. Heuristic searches that performed rearrangements on the original tree obtained by stepwise addition of the taxa yielded comparable values of support for bootstrap and jackknife. However, the fastest technique could reduce the time of the analyses by 30-fold.

These classical phylogenetic analyses are based on biological characters, such as morphological traits or DNA sequences, but supertree reconstruction methods have also been developed to build large phylogenetic trees by gathering the information directly from existing ‘source’ trees. An overlap of taxa between the source trees is sufficient for the methods to be applied, and the process allows very large trees to be created quickly. Several methods have been proposed to build supertrees and chapter 4 examined the ‘matrix representation using parsimony’ method. An empirical assessment using several different data sets from the grass family was made by comparing several modifications of this method. The data sets were analysed separately and the resulting topologies were used as source trees in the supertree reconstructions. Modifications that took into account the level of support present in the source trees produced supertrees that were closer to a classical analysis combining the different DNA sequences. Supertrees were also built from 55 published topologies for the grass family to create the largest grass phylogenetic trees containing 401 genera. The supertrees obtained highlighted interesting questions concerning the evolutionary history of the grass family, and the relationships between the clade comprising maize, wheat, and rice were further investigated in chapter 5. In this chapter, extensive simulations were performed to investigate whether the discrepancies between topologies obtained from different molecular data sets could be affected by random or systematic errors. The results indicated that several DNA sequences have a strong bias towards a particular placement of wheat. However, the general result suggested that the level of taxa and character sampling in studies of grass phylogenetics have not been sufficient to avoid high rates of errors and that these have impaired the ability of methods to correctly reconstruct grass evolutionary history. Finally, in response to the previous results, a large phylogenetic analysis of the *trnLF* and *rbcL* plastid regions is presented in chapter 6. The *rbcL* data set placed wheat as sister to maize, while this topology was only obtained with *trnLF* when Bayesian analysis was performed. With this DNA region, maximum parsimony analysis placed wheat within the BEP clade. The main subfamilies were supported, but the relationships between these groups could not be clearly defined. Divergence times were estimated by calibrating these phylogenetic trees with four grass fossils, suggesting a rapid diversification of the

grasses between 40 to 30 Mya. The calibrated dates also allowed an estimate of the appearance of the C_4 photosynthetic pathway in the grasses at 20 to 10 Mya, an origin that corresponded to low levels of past CO_2 concentrations. Therefore, CO_2 levels could have been a factor in the origin of C_4 photosynthesis in grasses, an adaptation that could have helped the huge diversification of this important angiosperm family.

CHAPTER 1. INTRODUCTION

Dobhzansky (1973) said that nothing in biology makes sense except in the light of evolution. A possible corollary stipulates that nothing in evolution makes sense except in the light of phylogeny. Although ' phylogeny' could have been replaced by many other fields in biology in the previous sentence, phylogenetic reconstructions are core components in our understanding of evolution. The term phylogeny means, in its etymology, simply the ' origin or birth of tribes' (Sporne, 1974 p. 167), but a phylogenetic tree is generally used as an hypothesis about the evolutionary history of a set of organisms, and provides a graphical estimate of the shared ancestry between organisms. As a consequence, all the events of biological evolution are played out somewhere along the branches of phylogenetic trees, where traces of the historical evolutionary processes that gave rise to the diversity of contemporary species are preserved. The recent revolution in DNA sequencing technology (Maxam and Gilbert, 1977; Sanger et al., 1977; Mullis and Faloona, 1987) have resulted in the production of more accurate phylogenetic trees, or gene genealogies, that can be used to help understand biological processes occurring at many different levels of life' s hierarchy. Scientists working in fields as different as behaviour, conservation biology development, ecology, epidemiology, evolution genetics, and are now united by common knowledge, methods, theories, and phylogenetic information is the glue tightening all this together (Harvey and Nee, 1996).

This thesis aims to investigate and advance phylogenetic methods and to improve phylogenetic understanding of the angiosperms in general and the grass family in particular. This chapter aims to cover the relevant issues in phylogenetics and introduce the taxonomic groups under investigation.

1.1 Historical overview of phylogenetics

Initial efforts to reconstruct phylogenetic history were based on few (if any) objective criteria, and estimates of phylogeny were little more than plausible assertions by experts on particular taxonomic groups. For example, phylogenetic hypotheses made for higher plants were often in the form of bubble diagrams or minimum spanning trees (e.g. Bessey, 1915; Sporne, 1956; Cronquist, 1981; Dahlgren et al., 1985), but groupings were based on overall morphological similarities and the presence of putatively primitive or advanced characters, and lacked objective methods for their construction. This led to Cain' s statement (1959, p. 243) that “young taxonomists are trained like performing monkeys, almost wholly by imitation." The situation gradually changed during the 1930s, 1940s, and 1950s, through the efforts of individuals like the botanist Walter Zimmermann (1934, 1943) and the zoologist Willi Hennig (1950, 1966). They began to define objective methods for reconstructing evolutionary history based on the shared attributes of extant and fossil organisms. In the 1960s, these methods were refined and developed into explicit criteria for estimating phylogenetic trees. Numerical methods were applied to systematics (Sokal and Sneath, 1963) and numerical inference of phylogenies led to the description of maximum parsimony (MP) and minimum evolution (ME) distance methods (Edwards and Cavalli-Sforza, 1963, 1964) for continuous variables. In an attempt to choose between these two approaches, Edwards and Cavalli-Sforza (1964) turned towards maximum likelihood (ML), which proved to be different from both of the methods. At about the same time, Camin and Sokal (1965) began using parsimony on discrete characters. Parsimony criterion was also soon used on protein sequence data (Eck and Dayhoff, 1966) and detailed descriptions of distance methods and their application to protein sequences were published (Fitch and Margoliash, 1967). The next step was to adapt the MP criterion to DNA sequences in order to cope for the degeneracy of the genetic code, a step that generated a certain amount of work (see for example Fitch, 1971; Moore et al., 1973; Fitch, 1974; Fitch and Farris, 1974; Moore, 1974; Moore, 1977). Since then, improvements in MP methods have mainly stayed within an algorithmic framework. Indeed, the criteria described during the 1970s have laid the basis for all implementations of MP methods, whereas the algorithms for estimating minimum-length trees are still being modified and improved (e.g. Swofford and Maddison, 1987; Maddison, 1989; Nanney et al., 1989; Wheeler and Nixon, 1995; Farris et al., 1996; Ronquist, 1996; Nixon, 1999). Other implementations of MP methods have been investigated more recently, including the estimation of ancestral character states (Maddison, 1991; Maddison, 1995; Martins and Hansen, 1997) or the use of MP on continuous characters (Rogers, 1986; Swofford and Berlocher, 1987; Huey and Bennet, 1987; Thiele, 1993; Wiens, 1995; Wiens, 2001).

MP has been the favoured method for inferring evolutionary trees, due mainly to its computational speed, its mathematical simplicity, and its apparent lack of assumption involving underlying models (Steel and Penny, 2000). However, the last decade has seen an increased tendency towards more statistical phylogenetics

(Felsenstein, 2001), and model-based approaches have come to rival MP for phylogenetic methodology. Distance methods can essentially be viewed as approximations to a full ML approach, because although the same sorts of models as ML are used to correct the observed differences between DNA sequences, the parameters included in the models of DNA substitution cannot be directly estimated from the data with distance methods (Cavalli-Sforza and Edwards, 1967; Swofford et al., 1996). The ML method is an attractive alternative to MP in spite of its computational burden. Instead of asking under what biological conditions existing methods can be justified as statistical methods, ML has the advantage of directly implementing statistical methods within the phylogenetic inference by explicitly making use of stochastic models of DNA evolution (Felsenstein, 1982), therefore being much more statistically sound. However, workable ML estimations have been difficult to design to such a degree that early attempts had to fall back on a MP method (Edwards and Cavalli-Sforza, 1964) or least-squares pairwise method (Cavalli-Sforza and Edwards, 1967) as a less desirable alternative. The breakthrough came from Felsenstein (1973), who was able to eliminate the problematic calculation over ancestral nodes from the ML estimation. More importantly, he later developed a pruning algorithm that allows a vast increase in speed during the probability summation required for ML estimation (Felsenstein, 1981; Swofford et al., 1996). New developments have succeeded in further reducing the computational cost of ML (Olsen et al., 1994; Lewis, 1998; Rogers and Swofford, 1998; Salter and Pearl, 2001), and recent implementations of Bayesian analysis have allowed a huge improvement in computational time required during the estimation of the likelihood function through the use of Markov chain Monte Carlo (MCMC) algorithms to obtain posterior probabilities of parameters of interest (Rannala and Yang, 1996; Mau and Newton, 1997; Yang and Rannala, 1997; Larget and Simon, 1999; Huelsenbeck and Bollback, 2001). Although both Bayesian analysis and ML make use of the likelihood function, Bayesian analysis has the advantage to express the uncertainty in parameter estimation directly through the marginal posterior probability distribution resulting from the Markov chain process. Such assessment of uncertainty with ML can only be performed through the lengthy procedure of bootstrapping the original data matrix (Larget and Simon, 1999).

A critical element in a ML or Bayesian estimation is to model how the probabilities of the various changes are calculated. The increasing availability of nucleotide and protein sequences has stimulated the development of stochastic models describing evolutionary change in molecular sequence over time (e.g. Jukes and Cantor, 1969; Kimura, 1980; Felsenstein, 1981; Hasegawa et al., 1985; Rodriguez et al., 1990; Tamura and Nei, 1993; Yang, 1993; Steel et al., 1993; Zharkikh, 1994). The most used molecular models have been designed as a homogeneous Markov process. However, all these models assume that the substitution rate of a position remains constant throughout time and that the sites are independent entities. More complex and certainly more realistic models involving the covarion hypothesis use a model where a site can switch between "on" and "off" states, which defines if the position can vary or not and therefore changes its rate of substitution along the tree (Fitch and Markowitch, 1970; Fitch and Ayala, 1994a,b; Miyamoto and Fitch, 1995; Lockhart et al., 1998; Galtier, 2001; Huelsenbeck, 2002; Lopez et al., 2002), limited degree of nonindependence between sites (von Haeseler and Schöniger, 1995) or variable selective pressure among sites (Nielsen and Yang, 1998; Yang et al., 2000; Yang and Swanson, 2002) are the subject of intensive work. Despite the early availability of a likelihood model for continuous traits (Felsenstein, 1973), the use of stochastic models has been restricted primarily to molecular data. Models have been applied to morphological traits, but the purpose of these models has been to infer ancestral states (e.g. Schluter et al., 1997; Mooers and Schluter, 1999; Pagel, 1999), to assess the magnitude of the evolutionary correlation between traits (Pagel, 1994), or to investigate the properties of other optimality criteria (Felsenstein, 1981; Goldman, 1990; Tuffley and Steel, 1997), and only one has been suggested for tree estimation (Lewis, 2001).

This historic summary does not cover the heated debates surrounding the early days of phylogenetics reconstruction either between proponents of phenetics versus cladistics or later between statistical phylogenetics versus cladistic parsimony. Fortunately and as stated by Felsenstein (2001, p. 467), "In the era of Markov chain Monte Carlo methods, Hidden Markov Models, and mathematical genomics, little sign is left of the harrowing conflicts of the 1980s," and it is maybe better that way. Biologists using phylogenetic inference nowadays think of phylogenetics as being basically statistical, and pragmatism has become the rule in their choices of methods rather than simple prior philosophical commitment (Felsenstein, 2001). As computers continue to get faster, and given the increasing availability of many software implementing more and more complex methods, it is more than likely that statistical phylogenetics will continue to increase its importance in our attempt to recover the evolutionary history of a set of organisms. However, the ease in obtaining molecular data is at the moment outpacing the increase in computer performance as well as our ability to develop better algorithms, therefore quick methods of phylogenetic reconstruction like MP will still play an important role in the near future, especially when large amounts of data will be gathered for analysis. Chapters 2, 3, and 4 of this thesis explore such issues.

1.2 Beyond classification purposes

Since 1981, the number of articles reporting phylogenies based on DNA sequence information has been increasing exponentially (Pagel, 1999) with 2064 published papers making reference to phylogenetics in 2001 and 1156 for the first six months of 2002 (ISI, Web of Science; personal observation). The prospect of describing in detail the patterns of descent within many of the major groups of organisms is now becoming a reality, and this has triggered a review of most taxonomic descriptions previously based on morphological characters (e.g. APG, 1998; GPWG, 2001). However, the influence of these new phylogenies extends beyond cataloguing the relatedness of species, and studies of phylogeny have permeated almost every subdiscipline in biology (Moritz and Hillis, 1996). This section therefore gives an overview of the wide range of possibilities offered by phylogenetic trees to investigate biological processes.

1.2.1 Ancestral character states

During phylogenetic inference, the states of characters represented at each node are usually not estimated or given and the end goal of the method is to obtain a topology and sometimes branch lengths (Swofford et al., 1996). But once a topology is estimated, ancestral states of characters mapped onto the tree can be a valuable source of information to describe what the past organisms were like, to discover how traits evolve and to better understand their function (Cunningham et al., 1998; Pagel, 1999). In molecular biology, reconstruction of ancestral proteins or genes could provide insight into how genes will respond when subjected to forced evolution methods, or be used to detect directional trends (Ivics et al., 1997; Golding and Dean, 1998). For instance, reconstructions of ancient artiodactyl ribonucleases gave insights into the evolution of enzymatic process involving the true ruminant digestion (Jermann et al., 1995), and estimation of GC content of the common ancestor to life suggested that it was not compatible with the high GC content that would indicate a thermophilic common ancestor to life (Galtier et al., 1999). Likelihood approach has been used to identify specific sites in plant chitinases that have been subject to selection (Nielsen and Yang, 1998), suggesting diversifying selection as a mechanism of defence against fungal pathogens in the mouse-ear cress (Bishop et al., 2000). Similar approaches have also been used on morphological characters to study sexual selection, and habit and diet preferences in birds for example (Schluter et al., 1997).

1.2.2 Timing of evolutionary events

A phylogenetic tree is an evolutionary hypothesis about a set of organisms, and its nested structure provides a temporal framework for the order of appearance of the nodes. Tests of relative ordering of branching points have shown that the node order on morphological cladograms generally agree with the known fossil record for most groups of organisms (Benton, 1996). Very early in the development of molecular techniques, the discovery that molecular divergence is roughly correlated with divergence of time has suggested that genes and their protein products might evolve at rates constant enough to be able to calibrate a molecular clock (Zuckerkandl and Pauling, 1962, 1965a,b). Although later developments viewed molecular differentiation not as a metronome, but as a Poisson process (Fitch, 1976; Wilson et al., 1977, 1987), which supported the idea of constancy of the underlying neutral mutation rate (Page and Holmes, 1998). However, heterogeneity of rates across different nucleotide positions, different genes, different genomic regions, or different genomes within an organismal lineage are undeniable (e.g. Li and Graur, 1991) and research on molecular clocks has now centred on whether substitution rates are constant enough within genes across evolutionary time (Page and Holmes, 1998). Unfortunately, a Poisson model is often rejected with high confidence when species are considered in an evolutionary framework (Li and Bousquet, 1992; Muse and Weir, 1992; Tajima, 1993), and diverse solutions have been proposed to deal with this problem. One of them was to simply try to find the ‘right’ molecule, which evolution conforms to the expectancy of a molecular clock (e.g. Kumar and Hedges, 1998), or to keep only the taxa that passed the tests of constancy of rates (e.g. Takezaki et al., 1995). Other more desirable solutions have been to make an explicit statistical model that accounts for the overdispersion of the process, by considering either that the mean numbers of substitutions are identical in two lineages, but the variances are higher than the mean (i.e. Poisson distribution is not the model of DNA evolution; Gillespie and Langley, 1979; Cutler, 2000) or that the number of substitutions in a lineage is Poisson distributed, but the mean number varies among lineages (Hasegawa and Kishino, 1989; Lynch and Jarrell, 1993;

Sanderson, 1997, 2002; Thorne et al., 1998). Valuable insights on many aspects of evolutionary history can be gained by knowing divergence times for a particular group. In chapter 6 of this thesis, the aim has precisely been to use these most recent techniques (i.e. non-parametric rate smoothing (NPRS) and penalized likelihood (PL), Sanderson, 1997, 2002) to better understand the origin of the grasses, and what could have triggered their rapid radiation to become one of the largest and most ecologically and economically influential families of angiosperms.

1.2.3 Tempo and mode of evolution

Fascinating and fundamental questions such as: 'what are the causes of speciation?' 'how do rates of speciation vary?' or 'do any features, either of the organisms or environmental, correlate with speciation?' are difficult to answer because direct observation is usually impossible and adequate fossil records required to investigate these topics are often missing (Panchen, 1992). However, phylogenetic trees can yield insights into the tempo and mode of evolution through their shapes reflecting the processes that generated them (Raup et al., 1973; Slowinski and Guyer, 1989; Nee et al., 1994; Purvis, 1996). Estimating speciation rates and the departure from a constant speciation rate model either over time, among regions or among taxa have been central questions, and much discussion has centred around such methods (e.g. Guyer and Slowinski, 1991; Harvey et al., 1994; Nee et al., 1994; Purvis et al., 1995; Pagel, 1997; Barraclough et al., 1999; Pybus and Harvey, 2000; Nee, 2001). The causes of speciation are fundamental issues in the study of speciation, and the prospect of obtaining complete species-level phylogenies have revitalised this area of research (Barraclough and Nee, 2001). The role of geographical isolation (Brooks and McLennan, 1991; Chesser and Zink, 1994; Barraclough et al., 1998; Berlocher, 1998; Chan and Moore, 1999; Barraclough and Vogler, 2000; Coyne and Price, 2000), ecological shifts and diversification (Bush and Smith, 1998; Sato et al., 1999; Schluter, 2001), and key adaptations or innovations (e.g. Heilbuth, 2000; Smith, 2001) in promoting speciation are now being investigated using species-level phylogenetic trees and major advances should be seen over the next few years.

1.2.4 Comparative methods

Hypotheses about the adaptive significance of a trait have long been tested by making comparisons among species, but the dependence on correlational evidence and the assumption that extrinsic evolutionary processes were largely responsible for shaping the form of a phenotypic trait have been thoroughly discussed as weakness of non-phylogenetic comparative analyses (Baum and Larson, 1991; Harvey and Pagel, 1991; Lauder et al., 1993; Reeve and Sherman, 1993). Independent contrasts (Felsenstein, 1985b) was one of the first methods proposed to test for comparisons of continuous traits, and a rapid increase in proposed methods quickly followed. For example, phylogenetic least-squares regression is a general approach using regression techniques to relate two or more characters and where the phylogenetic component is incorporated in a complex error structure (Martins and Hansen, 1997), while the spatial autoregressive model partitions the variation in each trait into a phylogenetic (i.e. predicted) and specific effect (Cheverud et al., 1985; Rohlf, 2001). A similar approach, the phylogenetic mixed model (Lynch, 1990) draws an analogy with quantitative genetics to partition data into an overall mean (i.e. ancestral state at the root of the phylogeny) and heritable (i.e. passed on between taxa along the phylogeny) and nonheritable (e.g. phenotypic plasticity) components, while phylogenetic eigenvector regression (Diniz-Filho et al., 1998; Diniz-Filho, 2001) uses principal coordinate analysis to extract the most relevant eigenvectors of a phylogenetic distance matrix before performing a direct ordination method with the eigenvectors as predictors.

A series of comparative methods have also been proposed to deal with discrete characters. Maddison' s (1990)concentrated changes test, for example, is used to test the correlation between two characters and counts the number of changes occurring on the tree for two characters. It assumes that changes along any branch on the tree are equally likely. The same assumption holds for Pagel's (1994, 1997) test based on maximum likelihood ratio method. Alternative methods have been considered by Ridley (1983), Burt (1989), Grafen (1989), and Grafen and Ridley (1996, 1997). An idea of a method for dealing with discrete characters, not pursued during this thesis, is to use canonical correspondence analysis as the ordination method in a way similar to phylogenetic eigenvector analysis (Diniz-Filho et al., 1998).

1.2.5 Genomics and evolution of development

By the end of 2001, the complete sequences of 63 microbial genomes were available, and this number could exceed 200 by the end of 2002 (Mount, 2001). The genomes of mouse-ear cress, rice, *Caenorhabditis elegans*, fruit fly, and yeast are also complete, while the genome sequence of human, mouse, pufferfish, zebrafish and rat are almost complete. Such an amount of information will make possible the cataloguing of the diversity of genes, regulatory sequences, and intervening regions. However, a major aim of genomics research is to identify differences between genomes of species and evolutionary theory can provide a framework for understanding the relationship between sequences and changes in gene function (Goldstein and Harvey, 1999; Charlesworth et al., 2001). Evolutionary approaches are particularly relevant to the study of the evolution of genes families (Semple and Wolfe, 1999; Ruvinsky et al., 2000) where data on the evolutionary history of genomes are essential to estimate the rate and cause of gene turnover or to understand the evolutionary events following the appearance of new members of gene families by comparing related species (Charlesworth et al., 2001). Well-studied examples include the mammalian major histocompatibility complex (Nei et al., 1997), plant disease resistance genes (Michelmore and Meyers, 1998), or *Adh* genes (Small and Wendel, 2000), for example. Using phylogenetic trees, duplications of large genome segments, or entire genome duplication, can also be investigated and hypotheses concerning the size of genomes can be tested (Hughes, 1999; Postlethwait et al., 2000; Gu et al., 2002; McLysaght et al., 2002).

Phylogenetic trees can also be a great help in our attempt to explain how developmental processes and mechanisms produce changes in morphology and body plans. The current molecular-based view of organismal relationships can provide a framework, even when incomplete, within which comparative developmental data can begin to be interpreted. The Hox gene clusters are a good example to illustrate this aspect. Phylogenetics of the animal kingdom have helped infer that a Hox gene cluster existed in the last common ancestor of all extant bilaterians, and that all phyla descended from this common ancestor possess a Hox gene cluster, while lineages that split off earlier (e.g. cnidarians or sponges) do have sequences resembling some Hox genes, no physical clustering between them have been demonstrated (deRosa et al., 1999). Similar approach addressing the question that gene duplication may have been important for most developmental evolution seems to suggest that they have been neither necessary nor sufficient (Holland, 1999). Plant development has not been left aside, and the new phylogenetic tree available for the angiosperms for example has helped understand the role of structural gene complexes like MADS or APETALIA (see Soltis et al., 2002a for a review) involved in the flower development.

Although uncertainties in phylogenetic estimations can sometimes be taken into account in some of the applications described in this section (e.g. Housworth and Martins, 2001), an accurate estimate of phylogenetic relationships is nevertheless more often a requirement. Moreover, branch lengths estimates are very often of particular importance for many applications of phylogenetic reconstructions, and accurate estimation is therefore also required. Unfortunately, phylogenetics is conceptually one of the most difficult areas in biology, and inferring events millions of years ago will always be hard. Phylogenetic reconstructions have come a long way since their beginning, and we certainly have more accurate estimates now than ever before. However, there is still a tremendous amount of work required and new applications of phylogenetics are likely to unearth more problems and trigger more theoretical work.

1.3 Phylogenetic inference

A ‘method’ for inferring an evolutionary tree can be divided into three distinct parts: the choice of optimality criterion (e.g. ML, additive-tree distance methods and MP), the search strategy over the space of trees (i.e. exhaustive, branch and bound, or heuristic) and assumptions (explicit or implicit) about the model of evolution (Steel and Penny, 2000).

1.3.1 Choosing an optimality criterion

Much debate has surrounded the diverse optimality criteria to infer phylogenetic trees, and particularly whether they possess the quality of consistency or not. Felsenstein’s (1978) paper demonstrating that MP could be statistically inconsistent under conditions of inequality of rates and/or high rates of evolution, have lead to the designation of the infamous ‘Felsenstein zone’ where methods of phylogenetic reconstructions could fail to recover the true tree (e.g. Hendy and Penny, 1989; Hillis and Huelsenbeck, 1994; Kuhner and Felsenstein, 1994; Huelsenbeck, 1995). Advocates of MP have also designated a particular zone within the tree space where MP outperforms ML due to long-branch repulsion (‘Farris zone’; Siddall, 1998), but a recent study demonstrated that the advantage of MP was the consequence of bias in the method that allowed MP to find the correct tree with less data than ML (Swofford et al., 2001). Other reasons for statistical inconsistencies than can affect all the existing methods, for instance when the process of substitution exhibits nonstationarity across the tree (Lockhart et al., 1994), have also been explored. However, the concept of statistical consistency is always relative to the model of evolution selected and model-based methods such as ML and distance methods are not safe from positively misleading the tree-building process (Penny et al., 1992; Chang, 1996). This raises the important question of model selection, and how one can be sure that the model chosen is the best for the data at hand (e.g. Lockhart et al., 1996; Penny and Hasegawa, 1997; Huelsenbeck, 1998). All the more since a model using too many parameters will perfectly fit the data, but will loose the ability to discriminate between different underlying phylogenetic trees (Steel et al., 1994; Yang et al., 1995). However, consistency is an asymptotic property, where the probability of reconstructing the true tree converges to certainty as the sequence length tends to infinity. Phylogenetic methods are used with molecular data of pregiven length, and random (i.e. due strictly to a limited sample size used to make the estimate) and systematic (i.e. due to incorrect assumptions in the estimation method) errors are more likely to impede phylogenetic inference (Swofford et al., 1996; Sanderson et al., 2000). A study of the impact of these types of errors on our understanding of the relationship within the grass family is presented in chapter 5 of this thesis, and this chapter asks in particular whether wheat is sister to maize or to rice.

1.3.2 Searching the tree space

Phylogenetic reconstruction methods belong to a set of problems called NP-complete (i.e. nondeterministic polynomial time complete), a property of computational decision problems that can not be solved in polynomial time, and for which no efficient algorithms for their solution are yet known to exist (Aussiello et al., 1999). When dealing with more than 20 taxa, phylogeneticists have no choice but to rely on heuristic search strategies in order to find an optimal topology (e.g. Swofford, 1996; Page and Holmes, 1998). However, these ‘quick and dirty’ algorithms are far from foolproof and the search over the tree space can be trapped in local optima with the risk of missing entirely the optimum tree, although algorithms have been designed to reduce this risk (Maddison, 1991; Lewis, 1998; Charleston, 2001; Quicke et al., 2001), without removing it entirely. However, searches involving several hundreds of taxa are becoming more common (e.g. Chase et al., 1993; Källersjö et al., 1998; Miadlikowska and Lutzoni, 2000; Omilian and Taylor, 2001; chapter 2, 3, 4, 6 of this thesis) and such algorithms still imply huge computational burden on tree searches (e.g. Rice et al., 1997). With large numbers of taxa, it is also illusory, at the moment, to try statistical approaches of phylogenetic reconstructions, even with the simplest heuristic options. The development of Bayesian techniques and MCMC algorithms have raised the hope of a major increase in speed for model-based methods, but this quickness introduces problems of memory requirement, even with the best programming skills. As an example, MrBayes is described by its authors as a 'memory hog' (Huelsenbeck and Ronquist, 2001) that requires many hundreds of megabytes depending on the size of the data set. Of course, this should change with the fast increase of computer power, but at the moment, not all biologists can have access to high performance computers. One solution could be provided by recent technical advances in DNA sequencing because as shown by Hillis (1996) with an angiosperm tree of 228 taxa, accuracy of simple methods like MP improved greatly as the number of characters increased. Whether Hillis' (1996)simulations are a hit in the dark remains to be seen (Purvis and Quicke, 1997), and chapter 2 of this thesis tries to generalise these findings by investigating increasingly large angiosperm trees and different DNA sequences.

1.3.3 Meta-analysis approaches

Another solution for phylogenetic reconstructions could come from alternative tree-building methods. Recently the meta-analysis approaches of supertree methods have come to prominence and have been seen as potentially useful approaches to build large composite trees (Sanderson et al., 1998; Bininda-Emonds, 2000; Semple and Steel, 2000; Bininda-Emonds and Sanderson, 2001). Chapter 4 of this thesis investigates one of the supertree building approaches, using empirical data on the grass family to draw conclusions on its limits and advantages.

1.3.4 Confidence intervals

Once a topology is obtained, it is difficult to know how much confidence we can have in each subtree forming a phylogenetic tree. Mueller and Ayala (1982), Felsenstein (1985a) and Penny and Hendy (1985) proposed to use resampling techniques to estimate sampling variance and approximate the distribution of characters among taxa for any given data set of interest. Both bootstrap and jackknife (Efron and Tibshirani, 1993) have been used in phylogenetics, with bootstrapping being the much more commonly used technique (Swofford et al., 1996). Originally, percentages obtained by these techniques were considered as a measure of repeatability (i.e. probability that a specified internal branch would be found in an analysis of a new independent sample of characters; Felsenstein, 1985a), but more recently they have been interpreted as a measure of accuracy (i.e. probability that the specified branch is contained in the true tree; Felsenstein and Kishino, 1993). However, Hillis and Bull (1993) have shown that bootstrap proportions are unbiased but highly imprecise estimates of repeatability, and biased estimates of accuracy. An interesting question revealed by the development of Bayesian method is whether bootstrap percentages, when applied to ML inference, could be considered as similar to posterior probability Efron et al. (1996) looked at this issue and concluded that in a Bayesian sense, traditional bootstrap percentages can be thought of as reasonable assessments of error for the estimated tree, while two-level bootstrap algorithms are required for bootstrap percentages to represent hypothesis-testing confidence levels. The two-level bootstrap algorithm is intended to compensate the fact that the statistic estimated in phylogenetics (i.e. the tree topology) is not a continuous function of the data. Rather, the tree is constant within large regions of the data-space, and then changes discontinuously depending on characters sampled (Efron et al., 1996). Whatever the meaning we give to bootstrap (or jackknife) percentages, the type of heuristic search used to build the trees for each pseudoreplicate could have a large impact on the estimates we get. This aspect is particularly important with very large data sets, where computational difficulties are likely to bias the results. It is precisely this aspect that is investigated in chapter 3 for a large matrix of angiosperms.

1.3.5 Networks

A basic assumption implicitly made in many methods of phylogenetic reconstruction is that the evolutionary relationships among organisms are best represented by a tree. However, the actual evolutionary history may not be particularly tree-like. Occurrences of recombination events between genes, hybridisation events between lineages or lateral transfers (e.g. retro-transposition) will mislead any methods of phylogenetic inference that model reality with a tree (Swofford et al., 1996; Page and Holmes, 1998). In such cases, other methods designed to build networks should be used instead. There has been recently an increase of interest in that research area and many different methods are now available (e.g. Bandelt and Dress, 1992; Strimmer and Moulton, 2000; Xu, 2000; Makarenkov and Legendre, 2001; Strimmer et al., 2001; Legendre and Makarenkov, 2002).

1.3.6 Coalescent process

The most interesting stage in the establishment and divergence of a lineage can be considered as violating both the assumption of equilibrium required for population genetics studies and the assumption of isolation of lineages required by phylogenetic methods (excluding the networks). In this in-between phase, population history will be reflected in the form of shared ancestral polymorphisms that can be used to reconstruct gene genealogies, and the utility of such genealogies is based on the rigorous conceptual framework of coalescent theory (Schaal and Leverich, 2001). The idea of the coalescent process is to consider the ancestral history of genes in a sample by developing a model for the time to common ancestry (Kingman, 1982), and coalescent theory tells us what gene genealogies are expected to look like if populations have different demographic histories by linking the distribution of divergence times among individuals with effective population size using the genetic diversity parameter $\theta = 4N_e\mu$ (Page and Holmes, 1998; Emerson et al., 2001). The integration of coalescent theory into a statistical framework has led to the growing development of coalescent-based methods to analyse genetic data. Several approaches have been used, such as ML estimates incorporating MCMC to integrate across several phylogenies (Kuhner et al., 1995; Kuhner et al., 1998; Beerli and Felsenstein, 2001), recursive evaluation of trees with fewer coalescence events using likelihood surfaces calculations (Bahlo and Griffiths, 2000), or information from a single tree topology and correcting the estimate through simulation (Fu, 1994).

As a conclusion, with the recent development of phylogenetics and the wealth of methods currently available, it is good to remember Felsenstein’s (1983, p. 331) words: "No method has a monopoly of virtue - each has value to the extent that its assumptions correspond to the biologist' s conclusions about the data."

1.4 Angiosperm and grass phylogeny

Both the angiosperms and one of their families, the grasses, have been used throughout this thesis as case studies for investigating different aspects of phylogenetic reconstructions, and this section gives a brief overview of the current understanding of their evolutionary histories. This introduction is not intended to be a detailed and exhaustive description of these two families.

1.4.1 Angiosperms

The angiosperms or flowering plants are the dominant group of land plants, and their monophyly is strongly supported in recent molecular studies (e.g. Chase et al., 1993; Doyle et al., 1994; Soltis et al., 1997; Savolainen et al., 2000; Qiu et al., 2000). However, the question whether the Gnetales represent the sister lineage to the angiosperms is still debated (e.g. Bowe et al., 2000; Chaw et al., 2000; Qiu et al., 2000; Sanderson et al., 2000; Friedman and Floyd, 2001; Rydin et al., 2002).

The recent molecular studies have reshaped our understanding of angiosperm systematics, and the traditional division of angiosperms into monocots and dicots does not hold since the arrival of new molecular data sets (Chase et al., 1993; Doyle et al., 1994; Soltis et al., 1997; Savolainen et al., 2000; Qiu et al., 2000). The monocots are still considered as a monophyletic group supported by the synapomorphies of leaves with parallel venation, embryo with a single cotyledon, sieve cell plastids with several cuneate protein crystals, stems with scattered vascular bundles, and an adventitious root system (Judd et al., 1999), as well as by 18S rDNA, *atpB*, *rbcL* and *matK* DNA sequences (Chase et al., 1993; Soltis et al., 1997; Qiu et al., 2000). However, in all published molecular analyses, the dicots form a paraphyletic complex, and the angiosperms have been split into three new groups: the monophyletic eudicots containing previous dicots taxa having tricolpate pollen; the Magnoliids including Magnoliales, Laurales and Illiciales that are woody plants with pollen grain lacking columellar exine structure; and the paleoherbs including the monocots, Aristolochiales, Ceratophyllales, Nymphaeales, and Piperales that are usually herbaceous plants with pollen grain having columellar exine structure (APG, 1998; Judd et al., 1999). However, except for the monophyletic eudicots and monocots, no clear and definitive division of the angiosperm has been proposed (APG, 1998). The relationship between these three groups is therefore still under investigation. Different topologies have been obtained depending on the data analysed (e.g. Savolainen et al., 2000; Qiu et al., 2000), and the rooting of the angiosperm is still equivocal (e.g. Qiu et al., 2001).

1.4.2 Poaceae

The grass family or Poaceae is with 10,000 species and 650 genera one of the largest angiosperm families (Mabberley, 1993). Grasses play a major role in human sustenance, either as cereal crops or as a source of forage, making them the focus of intensive scientific study, and the new development of genomics has attracted interest due to synteny of grass genomes (Keller and Feuillet, 2000).

The grasses have long been recognised as a natural group within the order Poales and the sister-group relationship between Poaceae and Joinvilleaceae has been repeatedly supported (Clark et al., 1995; Soreng and Davis, 1998; GPWG, 2001). Many comprehensive classifications of the family have been proposed (e.g. Stebbins, 1956; Clayton and Renvoize, 1986; Watson and Dallwitz, 1992; GPWG, 2001), and the grasses have recently been split into two clades: the BEP clade containing the Bambusoideae (e.g. bamboos), Ehrhartoideae (e.g. rice) and Pooideae (e.g. wheat, rye, barley); and the PACCAD clade containing the Panicoideae (e.g. maize, sugarcane), Arundinoideae, Centothecoideae, Chloridoideae (e.g. finger-millet), Aristidoideae and Danthonioideae (GPWG, 2001). A recent combined analysis of morphological characters, plastid restriction sites and six DNA sequences have started to reshape the subfamilial relationships within the grasses (GPWG, 2001). A small set of taxa (*Anomochloa*, *Streptochaeta*, *Guaduella*, *Puelia* and Phareae) has been represented as a varying number of early-diverging lineages within the Poaceae (Clark et al., 1995; Duvall and Morton, 1996; Soreng and Davis, 1998; Hsiao et al., 1999; Clark et al., 2000). The PACCAD clade has been found to be a strongly supported monophyletic group, with Panicoideae and Chloridoideae being monophyletic subfamilies within the PACCAD clade. However, the relationships within the PACCAD clade are still not well defined (Clark et al., 1995; Duvall and Morton, 1996; Soreng and Davis, 1998; Hsiao et al., 1999; Clark et al., 2000). The BEP clade however is still an area of controversy. The core Pooideae and the Bambusoideae *s.s.* are considered as strong monophyletic groups (Soreng et al., 1990; Davis and Soreng, 1993; Clark et al., 1995; Soreng and David, 1998; Clark et al., 2000), but the existence of a BEP clade has been challenged by several molecular studies (Duvall and Morton, 1996; Gaut et al., 1997; Barker et al., 1999; Gaut et al., 1999; Hsiao et al., 1999). The issue whether the Pooideae are sister to the PACCAD clade or whether they belong to the BEP clade is therefore still open, and the answer to this question will represent a major step forward in our understanding of the evolutionary history of the Poaceae.

1.5 Aims of this thesis

The general aim of this thesis was to study the effects of using large molecular data sets to reconstruct phylogenetic trees. One aspect was to understand how actual methods could handle data sets containing hundreds or thousands of taxa, and to investigate the effects of large data sets on the confidence levels obtained for phylogenetic inference (chapter 2 and 3) while another aspect was to empirically test a new method to build large composite phylogenetic trees (chapter 4). The two remaining parts focused on the grass family by first investigating the effect of random and systematic error on our estimation of grass phylogenetics (chapter 5), and secondly by using one of the largest grass data sets to estimate the origins of this important family of angiosperms (chapter 6). Specific objectives for each chapter are as follows:

Chapter 2:

- to assess whether large phylogenetic trees can be accurately reconstructed.
- to compare the performances of MP and NJ reached with *rbcL*, *atpB,* and the 18S rDNA to reconstruct large angiosperm phylogenetic trees.
- to investigate the feasibility of accurately reconstructing complete generic-level phylogenies for the angiosperms.

Chapter 3:

- to compare the bootstrap and jackknife percentages obtained with various heuristic searches using several of the swapping algorithms available.
- to provide some guidelines when assessments of support for large matrices are being carried out.

Chapter 4:

- to assess the relative merits of the supertree approach using the grass family as a case study.
- to investigate the effects of irreversible characters on MRP supertree reconstructions
- to evaluate the differences between the supertrees we obtained and an approach using combined data.

Chapter 5:

- to investigate the possible occurrence of error, bias, and inconsistency in grass phylogenetic reconstructions based on six different DNA regions.
- to assess whether random or systematic error could explain the disparate but strongly supported phylogenetic hypotheses that are obtained for the grass

family with different genes, or if it is necessary to seek other explanations for conflict between these molecular data.

Chapter 6:

- to increase the sampling of taxa in order to have a more comprehensive understanding of the relationships within the family.
- to provide evidence based on the *trnLF* region and an extended taxa sampling for the *rbcL* region.
- to date the divergence events within the family, and to examine the effects of levels of past CO_2 concentration on the origin of the C_4 photosynthetic pathway.

1.6 Structure of the thesis

Three papers taken from different chapters of this thesis have been already published or submitted in peer-reviewed journals. Chapter 4 is the basis of a paper published in *Systematic Biology* (Salamin et al., 2002) with Trevor R. Hodkinson and Vincent Savolainen as co-authors, while Chapter 3 is the basis of a paper in press in *Molecular Phylogenetics and Evolution* with Mark W. Chase, Trevor R. Hodkinson, and Vincent Savolainen as co-authors. A paper taken from chapter 5 has been submitted to *Proceedings of the Royal Society London, Series B* with Trevor R. Hodkinson and Vincent Savolainen as co-authors. Finally, a paper based on the results of chapter 2 is in preparation for submission to *Systematic Biology*, while the paper based on chapter 6 is in progress and will probably be submitted to the *Proceedings of the National Academy of Science, USA*. Each chapter is mutually exclusive, and none of them use the same technique or even the same approach to answer their specific questions. Therefore, in my opinion, it would not have been judicious to write a common Material and Methods, Results and Discussion chapter, and such layout would have rendered the thesis difficult to read and follow.

Chapter 8, called technical notes, was added in order to describe the software and other bioinformatic tools developed during the course of the thesis. These do not represent scientific questions, and as such, were difficult to integrate within the other chapters. However, a large amount of work has been spent on developing these tools, and without them it would has been difficult to obtain or analyse the results presented in this thesis. As such, they are an integral part of my thesis.

Finally, collaboration with my two supervisors on research outside the aim of my thesis has led to two additional publications. A study of the information content of coding plastid DNA sequences in the angiosperms have been published in volume 51(4) of *Systematic Biology* (Savolainen et al., 2002), while a study of phylogenetics of *Miscanthus*, *Saccharum* and related genera in *Journal of Plant Research* (Hodkinson et al., 2002b). Several other publications concerning the phylogenetics of the grass family are in preparation.

Chapter 2. Towards the Reconstruction of Complete Generic-Level Angiosperm Phylogenies: Are the Big Indeed Easy?

2.1 Introduction

A major challenge for systematists over the next decades is to assemble a 'Tree of Life', and both the European Union and the US National Science Foundation have supported discussion of making the reconstruction of such a tree a major research focus (Soltis and Soltis, 2001; V. Savolainen, pers. comm.). Complete or near-complete family-level phylogenies have already been built for the angiosperms (Savolainen, 2000; Qiu et al., 2000) and the next goal will be to obtain complete generic-level phylogenies for the flowering plants, a task that will involve dealing with around 13,000 taxa (Mabberley, 1993). Regardless of the scale of the phylogenetic problem, sampling of large numbers of taxa is a requirement in order to get a better understanding of macroevolutionary processes affecting a particular clade and to resolve systematic issues concerning the taxa in question. For many groups of organisms, this means sampling several hundreds or thousands of taxa. For example, the grasses have approximately 650 genera and 10,000 species and any meaningful sampling of the family would have to include hundreds of taxa. Moreover, with the advances of molecular techniques, large numbers of DNA sequences are being produced and more and more comprehensive phylogenetic trees are being analysed (e.g. Chase et al., 1993; Källersjö et al., 1998; Miadlikowska and Lutzoni, 2000; Omilian and Taylor, 2001). It is therefore important to understand whether actual methods of phylogenetic reconstructions are capable of accommodating large numbers of terminal taxa or whether new algorithms need to be developed (Soltis and Soltis, 2001).

2.1.1 Pitfalls of phylogenetics

Unfortunately, phylogenetic reconstructions belong to a set of computational decision problems that cannot be solved in polynomial time, and for which no efficient algorithms for their solution are known to exist (Aussiello et al., 1999). The difficulty being that the total number of possible trees increases more than exponentially with increasing number of terminal taxa (Felsenstein, 1978b; Steel, 1992). With a four-taxon case, Felsenstein (1978) showed that maximum parsimony (MP) could be inconsistent when rates of evolution along particular branches were high or when the difference in rate of evolution between branches was large. This case has since been generalized to multiple taxa (Hendy and Penny, 1989; Kim, 1996) and model-based methods (Chang, 1996), and a mathematical characterisation of the sufficient conditions for MP to be consistent have been found (Steel, 2001). Previous studies have also demonstrated that very large molecular data sets are often needed for accurate phylogenetic estimation. For instance, under extreme evolutionary rate variation, correct recovery of the phylogeny of just four taxa requires both an accurate DNA substitution model and information on tens of thousands, to millions, of nucleotides (Hillis et al., 1994).

Although dealing with large DNA matrices could, thus, seem a priori an impossible task, Hillis (1996) reached an opposite conclusion. Using computer simulations, he showed that MP and neighbour-joining (NJ) could easily retrieve a model tree based on the 18S large subunit of nuclear ribosomal DNA (18S rDNA) for 228 angiosperm taxa from Soltis et al. (1997). Low numbers of characters (from 3,000 bp) were needed for MP and NJ to converge towards 100% of the tree nodes correctly resolved (Hillis, 1996). Graybeal (1998) investigated the effect of increasing either the number of taxa or the number of characters in a phylogenetic analysis. Starting from the four-taxa tree typical of the ‘Felsenstein zone’ (Felsenstein, 1978a), the accuracy of the phylogenetic reconstruction was improved dramatically with the addition of taxa, whereas the improvement in accuracy was much less perceptible when the numbers of characters were increased. However, the way the taxa are added to a growing tree has also been shown to have a large impact on the accuracy (Kim, 1998). There is a requirement that the long branches have to be broken by adding taxa to see an improvement in accuracy. Furthermore, using a birth-death process with taxon sampling to model cladogenesis, Rannala et al. (1998) reached the conclusion that for a given number of taxa, the accuracy of the inferred phylogeny is increased if the terminal taxa represent a more complete sample of the extant taxa. Simply including more taxa will not increase the accuracy of the inferred phylogeny if they are poorly sampled of if these additional taxa share a more distant ancestor than the ingroup. Purvis and Quicke (1997) proposed that the 18S rDNA tree for the angiosperms used by Hillis (1996) could be considered as a ‘perfect’ tree, which by chance is well suited for parsimony analysis (i.e. the mean number of substitutions per site in the tree is low, which will suit MP). However, MP easily retrieved the same 18S rDNA phylogenetic tree even with a ten-fold increase in the expected number of substitutions (Purvis and Quick, 1997; Hillis, 1998).

2.1.2 Aims

This chapter aims to assess whether reconstructions such as Hillis' (1996) 18S rDNA tree are exceptions or whether large phylogenetic trees can often be accurately reconstructed. In other words, are the big indeed easy? Angiosperm phylogenies can help investigate this problem, because a number of large DNA matrices containing several hundreds of taxa are now available. The same set of taxa form the backbone of each of these matrices, with the largest matrix containing 567 taxa representing almost all angiosperm families. Moreover, 18S rDNA sequences are available for all the taxa of these matrices, allowing a direct examination of Hillis' (1996) results and an assessment of how good the dispersion of noise is when sampling density of taxa increases. The two plastid genes *rbcL* and *atpB* have also been sequenced for the taxa in each matrix and direct comparisons of performances reached with these two genes and the 18S rDNA can be made. In this chapter, the accuracy of MP and NJ analysis to reconstruct large phylogenetic trees was investigated by comparing the results obtained from simulations based on four different angiosperm DNA matrices containing 141, 228, 357 and 567 taxa. The results of Hillis (1996) were expanded by firstly investigating several angiosperm trees containing an increasing number of taxa, secondly using different model topologies for each simulated DNA matrix and thirdly considering the effect of two plastid genes and the 18S rDNA on MP and NJ performances. The feasibility of accurately reconstructing complete generic-level phylogenies for the angiosperms were also investigated by comparing the performances of MP and NJ to build trees containing 13,000 taxa.

2.2 Material and Methods

2.2.1 Data matrices and phylogenetic trees

The matrix used by Hillis (1996) containing 228 angiosperm taxa for 18S rDNA from Soltis et al. (1997) was reanalysed. Two other large matrices containing two plastid genes *atpB* and *rbcL*, as well as the 18S rDNA were also analysed, representing 141 (Chase and Cox, 1998) and 567 (Soltis et al., 2000) taxa in total. The fourth matrix used contained 357 taxa sequenced for *atpB* and *rbcL* (Savolainen et al., 2000). In order to have an 18S rDNA matrix of similar size to the 357 taxa one, a subset of 320 taxa was taken from the 567 taxa matrix that was also present in the 357 taxa matrix. These four matrices with 141, 228, 357 and 567 taxa have been previously analysed in their respective publications, and the most parsimonious trees published in these studies were kept as the reference trees to be used in subsequent simulations. For the matrix with 320 taxa and 18S rDNA, the 37 taxa that had no corresponding entry in the 567 taxa matrix were pruned from the most parsimonious tree published in Savolainen et al. (2000). All these trees are based on combined analyses of *atpB*, *rbcL*, and 18S rDNA (matrices with 141 and 567 taxa) or *atpB* and *rbcL* (matrix with 357 taxa). In order to make valid comparisons with Hillis' (1996) analyses, MP analyses were performed on 18S rDNA alone for each matrix (i.e. 141, 320 and 567 taxa). Heuristic searches with 100 replicates of random addition sequence were used keeping up to ten trees at each replicate, followed by nearest-neighbour interchange (NNI) swapping. One of the equally most parsimonious trees was kept as the model tree for further analyses. In total, four trees with 141, 228, 320 and 567 terminal taxa were obtained from 18S rDNA alone (hereafter 18S model trees), and three trees with 141, 357 and 567 terminal taxa were obtained from a combination of *atpB, rbcL* and 18S rDNA (hereafter combined model trees).

The branch lengths obtained by ML were estimated for all partitions of the four DNA matrices (i.e. 18S rDNA for the 18S and combined model trees, and *atpB*, *rbcL* and their three respective codon positions separately for the combined model trees) and a Gamma distribution was fitted on the branch length distributions using the technique of nonlinear regression analysis (Bates and Watts, 1988). The shape of the distributions was estimated using the software R version 1.4.0[1] by grouping the branch lengths into categories of 0.001 expected substitutions per site. This approach indicates the amount of heterogeneity that exists among branch lengths within a tree, but it does not suggest if short and long branches are intermixed along the tree, which could create conditions for MP to be inconsistent (Felsenstein, 1978a). To assess this, the ratio of length of each internal branch to its longest daughter branch was calculated, which could give an idea of the heterogeneity of lengths between adjacent pairs of branches. If the two branches are of similar lengths, the ratio will be close to one. A small parent branch and a long daughter branch will have a ratio lower than one, while a long parent branch and a small daughter branch will have a ratio larger than one. The ratios ρ were grouped as percentages into five categories: $\rho < 0.25$, $0.25 \leq \rho < 0.5$, $0.5 \leq \rho < 0.75$, $0.75 \leq \rho <$

[1] http://www.r-project.org

1 and $\rho \geq 1$. We also computed a statistic for the observed number of substitutions per site by summing the MP branch lengths on each model tree and dividing the total by the length of the actual sequence from each partition of the DNA matrices.

2.2.2 Simulations

In order to simulate DNA sequences, a model of DNA substitutions as well as a tree with branch lengths, representing the number of substitutions per site, are required. The HKY85+Γ substitution model (Hasegawa et al., 1985; Yang, 1994) was chosen to perform the simulations. First the parameters for the model were estimated using PAUP*4b (Swofford, 2000) under ML for each 18S tree based on 18S rDNA, and for each combined tree based on *atpB, rbcL*, their three codon positions and 18S rDNA. The transition to transversion ratio (ti/tv ratio) and the shape of the Gamma distribution for the rate of heterogeneity among sites were estimated from the DNA sequences, and the empirical base frequencies were used as an estimate of the equilibrium frequency of each nucleotide. These parameters were then used to estimate the branch lengths of each 18S and combined tree based on each data partition as described above under ML. We used the program evolver from the PAML3.1 package (Yang, 1997) to simulate matrices of different sizes based on each model tree. For MP, only one replicate was performed in each case because of the computational burden involved in handling such large matrices, while 20 replicates were analysed with NJ. For the simulations based on 18S rDNA using 18S and combined model trees and those based on *atpB* and *rbcL* using the combined model trees, DNA sequences of 100, 500, 1,000, 3,000, 5,000, 7,000 and 10,000 bp were created for 141, 228, 320 and 567 taxa. For the simulations of the three codon positions of *atpB* and *rbcL*, DNA sequences of 100, 1,000, 5,000, and 10,000 bp were created for 141, 357 and 567 taxa.

Each simulated data set was then subjected to MP or NJ analysis using PAUP*4b. For MP, heuristic searches were performed with 100 random replicates of stepwise addition followed by tree bisection and reconnection (TBR) swapping keeping only ten trees at each replicate. For NJ, K2P (Kimura, 1980) and the HKY85+Γ distances were calculated before building the tree. The percentages of tree correct (i.e. proportion of nodes correctly inferred) were calculated with the software TreeCorrect1.2[2] by comparing the nodes of model tree used to simulate the data with the nodes present in the most parsimonious tree(s) found. When

[2] see Chapter 8. Technical notes for its description

multiple equally most parsimonious trees were obtained, a node was considered as correct only if it was found in all trees saved. This measure is very conservative, and obtaining a correct node in one tree out of 100 could already be seen as a good result. However, when this approach was considered, the percentages of nodes correct obtained did not exceed substantially those obtained with the more conservative approach (data not shown). For NJ, the percentages were averaged over the 20 replicates performed, and the standard deviation was calculated.

2.2.3 A 13,000 taxa tree

Data sets containing 13,000 taxa were simulated using the parameters for the HKY85+Γ model (i.e. among site rate heterogeneity, ti/tv ratio and base frequencies) derived from the 18S rDNA for 567 taxa. A topology was first created using a Yule process (Steel and McKenzie, 2001) where each branch had an equal probability of generating the next lineage, until the required number of taxa had been reached, and a tree having an imbalance index (0.687; Fusco and Cronk, 1995) similar to the 567 taxa tree was selected using the software GenTree0.5[3]. Once the topology was created, branch lengths were assigned by randomly drawing branch lengths from a Gamma distribution having the same shape parameter (1.161; Table 2.2) as the one estimated from the 567 taxa matrix for 18S rDNA based on the 18S model tree. A modified version of evolver was used to create data sets containing 100, 500, 1,000, 3,000, 5,000, 10,000, 15,000, and 30,000 bp. The MP and NJ analyses were run on a 32 node IBM NetFinity cluster (each node with 2 x Intel Pentium III, 1000 MHz, 1GB RAM) available at the University of Dublin, Trinity College using PAUP*4b. Due to the computational time and memory required, only one replicate was performed for each data set, and MP analyses were performed using simple addition sequence followed by NNI swapping, while NJ analyses were performed using the HKY85+Γ distance. The resulting trees were then analysed using TreeCorrect1.2 as described above.

A logarithmic model was fitted onto the percentages obtained by MP using the linear regression module implemented in the software R version 1.4.0[4]. The theoretical curve obtained was then used to extrapolate the number of characters required to correctly infer 100% of the nodes from the 13,000 taxa tree.

[3] see Chapter 8. Technical notes for its description

[4] http://www.r-project.org

2.3 Results

2.3.1 Parameter estimation

ML estimation of the parameters for the ti/tv ratio, the percentage of GC content and the rate of heterogeneity among sites are shown in Table 2.1. The increase in number of taxa does not affect greatly the parameter estimates. Third codon positions of *rbcL* and *atpB* had the highest ti/tv ratio values ranging between 2.636 and 2.938 (Table 2.1), while second codon positions of *rbcL* had the lowest ti/tv ratios (0.588 to 0.598; Table 2.1). In contrast, second codon positions of *atpB* had much higher ti/tv ratios than *rbcL*. When all codon positions were considered together, more transitions than transversions were found for *atpB*, *rbcL* and 18S rDNA (1.812 to 2.555; Table 2.1), with generally *rbcL* having the lower and *atpB* the highest ratios.

The GC content was similar between the two coding genes (Table 2.1). Third codon positions of both coding genes deviated more from the parity AT/CG with only ca. 30% of GC (Table 2.1). Second codon positions had a GC content between 0.415 and 0.434, while first codon positions had more GCs than ATs in both *atpB* and *rbcL* (Table 2.1). All codon positions together resulted, for both *atpB* and *rbcL*, in GC content similar to the second codon positions; values that were closer to the 18S rDNA for 228 taxa than the other three 18S rDNA partitions (between 0.418 and 0.441 versus between 0.495 and 0.499; Table 2.1).

The last parameter estimated was the shape of the Gamma distribution of rate heterogeneity among sites. For most partitions, there was a decrease in heterogeneity among sites (increase of the shape parameter) when more taxa were added, although the differences were small (Table 2.1). The third codon positions had by far the highest values for shape of the distribution with values ranging from 0.932 to 1.382 depending on the coding gene, *atpB* having again higher values than *rbcL* (Table 2.1). First and second positions had more similar values, with more heterogeneity within the second codon positions. The values for 18S rDNA were similar to the first and second codon positions, while all three codon positions together had slightly higher values for the shape of the distribution (Table 2.1).

Table 2.1 - ML estimates of the parameters of the HKY85+Γ model of subsitution for the four size of trees used during the simulations.

Number of taxa	Sequences	ti/tv ratio	GC content	shape
141	*rbcL*	1.948	0.441	0.361
	rbcL 1	0.768	0.581	0.231
	rbcL 2	0.591	0.433	0.175
	rbcL 3	2.938	0.308	0.984
	atpB	2.555	0.428	0.348
	atpB 1	1.458	0.574	0.251
	atpB 2	2.154	0.417	0.192
	atpB 3	2.922	0.292	1.288
	18S rDNA	2.132	0.495	0.193
228	18S rDNA	2.170	0.421	0.228
357	*rbcL*	1.812	0.441	0.456
	rbcL 1	0.723	0.581	0.284
	rbcL 2	0.598	0.434	0.258
	rbcL 3	2.713	0.305	1.063
	atpB	2.337	0.427	0.472
	atpB 1	1.374	0.571	0.330
	atpB 2	1.778	0.415	0.253
	atpB 3	2.661	0.296	1.338
	18S rDNA	2.492[§]	0.498[§]	0.236[§]
567	*rbcL*	1.897	0.439	0.478
	rbcL 1	0.755	0.580	0.369
	rbcL 2	0.588	0.434	0.273
	rbcL 3	2.636	0.303	0.932
	atpB	2.424	0.427	0.491
	atpB 1	1.449	0.573	0.342
	atpB 2	1.831	0.418	0.257
	atpB 3	2.733	0.292	1.382
	18S rDNA	2.450	0.499	0.291

[§] parameters are estimated on the 320 taxa subtree instead of the 357 taxa tree

2.3.2 Branch length and number of substitutions

The shape of the Gamma distribution fitted on the branch length distributions for each DNA sequence is given in Table 2.2. The value of the shape parameter ranged from 0.878 to 1.161 for the 18S model tree branch lengths, with larger values for 357 and 567 taxa trees (Table 2.2). The values for the combined model tree with branch lengths estimated on 18S rDNA alone were slightly lower (0.871 to 1.060), while those estimated on *atpB* and *rbcL* were higher than for 18S rDNA (1.174 to 1.993; Table 2.2). When looking at the branch length distributions obtained from the codon positions of *atpB* and *rbcL*, first and second codon positions had lower values than the third codon position or the 18S (Table 2.2). Second codon positions for *atpB* and *rbcL* taken separately had values ranging from 0.307 to 0.471, first codon position values were between 0.415 and 0.578 and third codon position values were between 0.903 and 1.212 (Table 2.2).

The mean number of substitutions per site value increased with the addition of more taxa (Table 2.3), and values for 18S were similar to the second codon positions for *atpB* and *rbcL*. First codon positions had 1.5-2 fold more substitutions per site than the second codon positions, while the increase in the number of substitutions per site for the third codon positions was 7-9 fold in comparison to the second codon position (Table 2.3).

The ratios for lengths of each internal branch divided by its longest daughter branch are shown in Table 2.4. The values represent percentages of ratios found in each category. The first category of ratios ($\rho < 0.25$), where the parent branches were more than 4 times smaller than their daughter, represented the majority of ratios found in each topology and character partitions (Table 2.4). The percentages within this category varied from 38.62% for 18S rDNA on the 18S model tree containing 320 taxa to 73.18% for the second codon positions of *atpB* on the combined model tree containing 141 taxa (Table 2.4). A recurrent pattern was that second codon positions, for all tree sizes and for both coding genes, had ca. 70% of their ratios in this category. They were followed by the first codon positions that had between 55.31 and 66.66 % of their ratios in the first category, while third codon positions had similar percentages to *atpB*, *rbcL* and 18S rDNA (Table 2.4). The next categories with the highest percentage of ratios were the second ($0.25 \leq \rho < 0.5$) and fifth ($\rho \geq 1$) with similar percentages of ratios, while the third ($0.5 \leq \rho < 0.75$) and fourth ($0.75 \leq \rho < 1$) were the least represented categories (Table 2.4).

Table 2.2 - Estimates of the shape of the Gamma distribution (with a scale of 1) from the branch length distribution for the four tree sizes based on all data partitions.

		Number of taxa			
Topology	Sequence	141	228	357	567
Combined	*rbcL*	1.174	-	1.646	1.837
	rbcL 1	0.415	-	0.491	0.518
	rbcL 2	0.307	-	0.459	0.465
	rbcL 3	0.903	-	1.064	1.172
	atpB	1.431	-	1.761	1.993
	atpB 1	0.547	-	0.578	0.516
	atpB 2	0.362	-	0.440	0.471
	atpB 3	0.942	-	1.037	1.212
	18S rDNA	0.871	-	1.025§	1.060
18S	18S rDNA	0.919	0.878	0.945§	1.161

§Parameters were estimated on the pruned *atpB+rbcL* tree of Savolainen et al. (2000) containing 320 taxa instead of 357.

Table 2.3 - Statistics of mean number of substitutions per site for the four tree sizes based on all data partitions.

		Number of taxa			
Topology	Sequence	141	228	357	567
Combined	*rbcL*	4.389	-	8.996	13.741
	rbcL 1	2.233	-	4.983	7.424
	rbcL 2	1.216	-	2.721	4.103
	rbcL 3	9.515	-	19.286	28.411
	atpB	4.421	-	8.743	12.068
	atpB 1	1.852	-	3.829	5.170
	atpB 2	1.119	-	2.549	3.169
	atpB 3	10.293	-	19.851	27.809
	18S rDNA	1.784	-	3.118§	4.693
18S	18S rDNA	1.702	2.121	2.804§	4.502

§ Parameters were estimated on the pruned *atpB+rbcL* tree of Savolainen et al. (2000) containing 320 taxa instead of 357.

Table 2.4 - Ratios of each internal branch length divided by its longest daughter branch length grouped by categories.

Number of taxa	Topology	Sequence	$\rho <$ 0.25	$0.25 \leq \rho$ < 0.5	$0.5 \leq \rho$ < 0.75	$0.75 \leq$ $\rho < 1$	$\rho \geq 1$
141	combined	*rbcL*	47.48	20.86	14.38	7.19	10.09
		rbcL 1	64.49	10.86	5.79	5.79	13.07
		rbcL 2	71.73	12.31	4.38	5.79	5.79
		rbcL 3	47.82	22.46	8.69	9.44	11.59
		atpB	56.83	18.70	7.91	5.05	11.51
		atpB 1	66.66	11.59	5.79	4.37	11.59
		atpB 2	73.18	7.27	5.79	5.07	8.69
		atpB 3	55.79	18.84	9.46	5.07	10.86
		18S rDNA	56.52	16.66	11.59	2.89	12.34
	18S	18S rDNA	42.75	23.91	12.31	7.27	13.76
228	18S	18S rDNA	41.70	22.86	10.76	6.75	17.93
357	combined	*rbcL*	41.80	24.57	11.58	7.08	14.97
		rbcL 1	55.64	14.40	7.90	4.55	17.51
		rbcL 2	68.36	8.75	4.54	3.67	14.68
		rbcL 3	45.76	19.49	9.88	6.23	18.64
		atpB	49.71	14.40	10.45	7.08	18.36
		atpB 1	65.25	8.75	5.64	6.24	14.12
		atpB 2	70.90	9.03	1.41	5.39	13.27
		atpB 3	48.87	14.12	7.06	8.49	21.46
		18S rDNA	55.52[§]	14.51[§]	7.25[§]	4.11[§]	18.61[§]
	18S	18S rDNA	38.62[§]	22.22[§]	12.16[§]	6.37[§]	20.63[§]
567	combined	*rbcL*	45.56	19.32	11.87	6.77	16.48
		rbcL 1	55.31	14.71	9.57	4.07	16.34
		rbcL 2	69.68	10.28	3.19	3.36	13.49
		rbcL 3	47.87	15.78	10.28	4.25	21.82
		atpB	48.04	15.95	8.51	5.31	22.19
		atpB 1	63.47	9.57	4.07	6.41	16.48
		atpB 2	69.68	8.15	3.54	4.27	14.36
		atpB 3	48.22	15.42	8.86	4.07	23.43
		18S rDNA	48.75	18.26	8.86	4.78	19.35
	18S	18S rDNA	42.63	21.57	10.00	4.73	21.05

[§] Parameters were estimated on the pruned *atpB*+*rbcL* tree of Savolainen et al. (2000) containing 320 taxa instead of 357.

2.3.3 Simulations of 18S rDNA, atpB *and* rbcL

Efficiency of MP and NJ using HKY85+Γ distance, for estimating the model trees from the simulated data sets based on parameters estimated from the 18S rDNA are shown in Figure 2.1. For the simulations based on the combined model tree (Fig. 2.1A), MP and NJ methods correctly inferred less than 85% of nodes with 10,000 bp. Starting with 100 bp, there was a steep increase in the percentages of tree correct values found until the sequence length reached 1,000 bp. Then, a plateau was reached with each simulated data set from 1,000 or 3,000 bp for MP and NJ respectively, with a slow increase in percentages afterwards. This pattern was found in all simulations performed (Figs. 2.1 to 2.6). The large tree with 567 taxa proved more difficult to recover for both methods than the smaller 141 taxa tree with any sequence lengths, while the lowest percentages were found with the 320 taxa tree, from which less than 75% and 65% of the nodes were correctly placed with 10,000 bp for MP and NJ respectively (Fig. 2.1A). A different pattern appeared when one of the most parsimonious trees obtained for the 18S rDNA of each matrix was used as a model tree (Fig. 2.1B). With 10,000 bp, MP correctly inferred 99% of the nodes from the 141 taxa tree, with 97% reached with 3,000 bp (Fig. 2.1B). Slightly lower values were obtained with the 228 taxa tree, a result identical to Hillis' (1996) analyses (Fig. 2.1B). The two larger trees with 320 and 567 taxa obtained similar percentages that were lower than the two smaller trees (93% with 10,000 bp; Fig. 2.1B). The results obtained with NJ followed the same pattern to MP with a decrease in the percentages of tree correct values as the size of the tree increased (Fig. 2.1B), except that the percentages were slightly lower. This pattern was also found in Hillis' (1996) analyses.

The simulations performed on the combined model trees with parameters estimated from *atpB* and *rbcL* and analysed with MP and NJ using HKY85+Γ distance are shown in Figure 2.2. The percentages of nodes correctly inferred by both methods also quickly reached a plateau with only a slow increase in percentages as the sequence length increased from 3,000 bp. Simulations based on *atpB* resulted in MP and NJ recovering 90% of the nodes from the 357 and 567 taxa trees with 10,000 bp, but only 82% of the nodes from the smaller 141 taxa tree were correctly inferred with MP (Fig. 2.2A). Results obtained with simulated data sets based on *rbcL* were similar for both methods with the smaller 141 taxa tree being more difficult to infer (Fig. 2.2B). This was particularly true for the simulations analysed with NJ that showed extremely large standard deviation values with the 141 taxa tree, indicating large variation between the simulated data sets (Fig. 2.2A).

The simulated data sets were also analysed by NJ using the simpler K2P distance, without taking into account the rate of heterogeneity among sites (Fig. 2.3). Although the relative pattern found remained similar to NJ using the HKY85+Γ distance for each DNA sequence or model trees analysed, the percentages of nodes correctly inferred were constantly lower. The differences in percentages of nodes correctly recovered between the two distances used ranged from a few percentages with 18S rDNA simulations based on 18S model trees to more than 10% with 18S rDNA simulations based on combined model trees (Fig. 2.3 vs. Figs. 2.1 and 2.2).

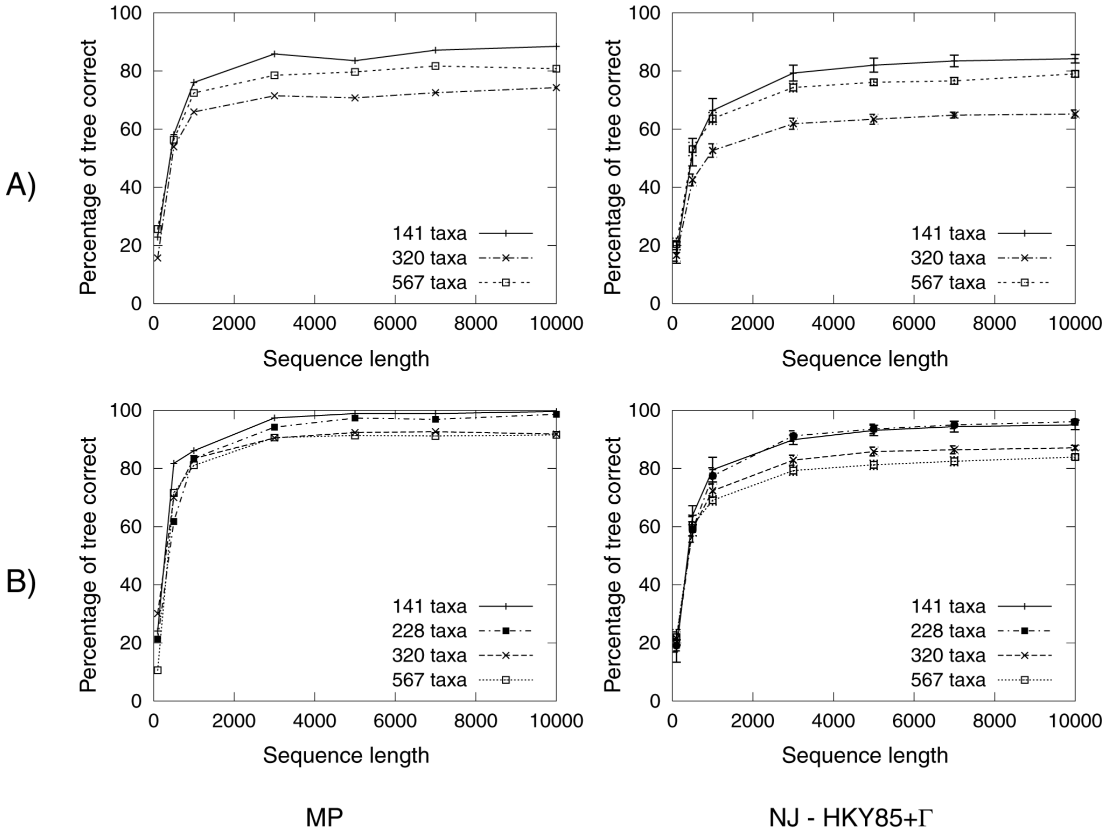


Figure 2.1 - Efficiency of MP and NJ (HKY85+Γ distance) for estimating the different model trees from the simulated data sets based on 18S rDNA parameters. A) Combined trees based on combined analyses of atpB, rbcL *and 18S rDNA (matrix with 141 and 357 taxa) and on combined analysis of atpB and rbcL (matrix with 320 taxa). B) 18S trees based on analyses of 18S rDNA for 141, 228, 320 and 567 taxa. For NJ, vertical bars represents standard deviation around the mean.*

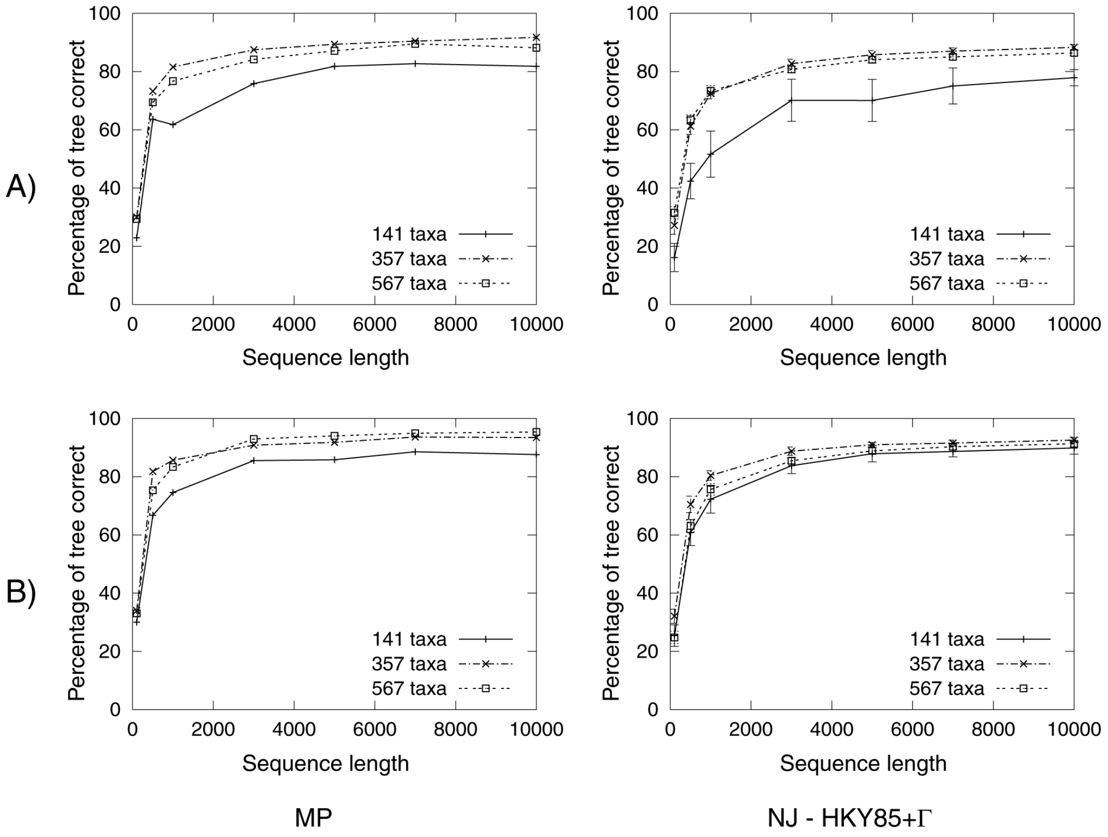


Figure 2.2 - Efficiency of MP and NJ (HKY85+Γ distance) for estimating the combined model tree from the simulated data sets from 141, 357 and 567 taxa. A) Simulations based on atpB *parameters. B) Simulations based on* rbcL *parameters. For NJ, vertical bars represents standard deviation around the mean.*

Figure 2.3 - Efficiency of NJ algorithm using K2P distances for estimating the different model trees from the simulated data sets. Vertical bars represents standard deviation around the mean. A) Simulations based on 18S rDNA parameters estimated on the combined model tree. B) Simulations based on 18S rDNA parameters estimated on the 18S model tree. C) Simulations based on atpB *parameters estimated on the combined model tree. D) Simulations based on* rbcL *parameters estimated on the combined model tree.*

(see figure next page)

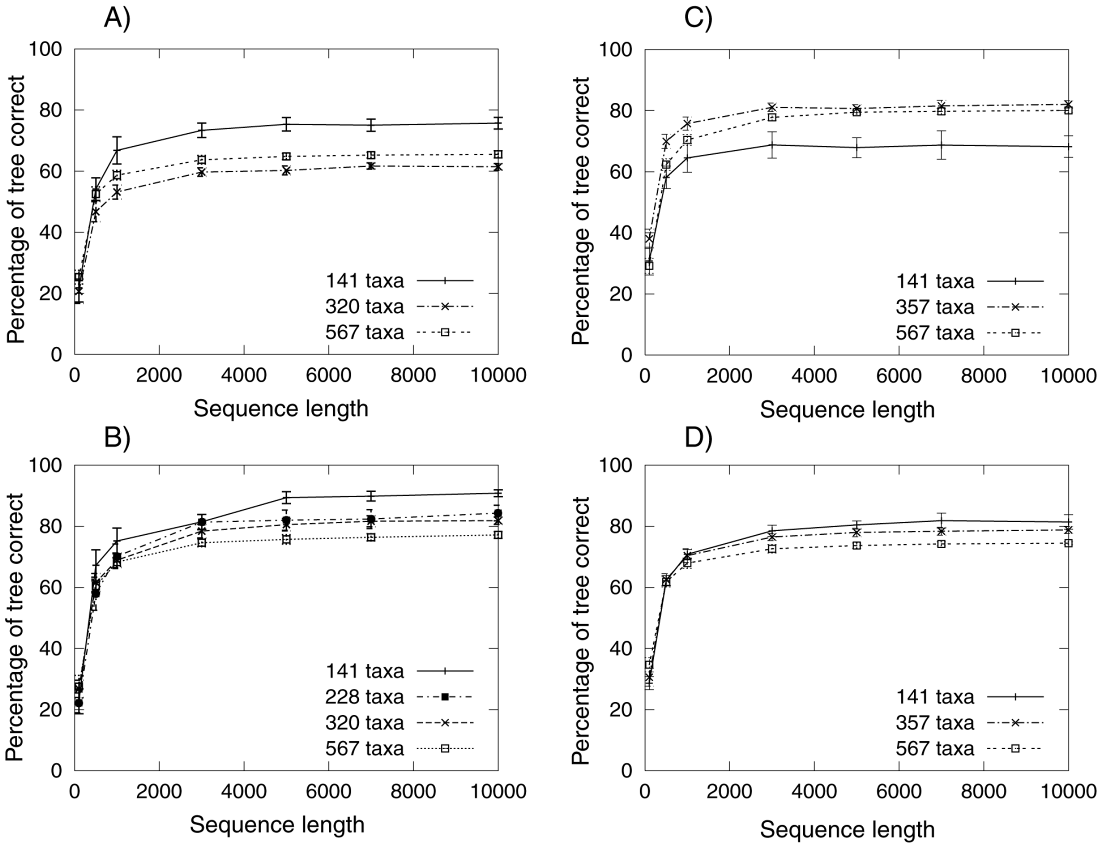


Figure 2.3

2.3.4 Simulations based on three codon positions

The results of the simulations based on the three codon positions of *atpB* and *rbcL* are shown in Figures 2.4 and 2.5. Simulations analysed with NJ using HKY85+Γ distances (Fig. 2.4) gave similar results to simulations analysed with MP (Fig. 2.5) with a plateau reached in a similar manner as in the previous simulations. With both methods, and for the three tree sizes under investigation, simulated data sets based on the second codon positions were very difficult to reconstruct with only 40% of the nodes correctly inferred for *atpB* and just 43% for *rbcL* with 10,000 bp (Figs. 2.4B and 2.5B). Simulations based on the first codon positions performed slightly better, reaching between 50 and 55% for *atpB* and between 50 and 64% of nodes correctly inferred for *rbcL* again with 10,000 bp (Figs. 2.4A and 2.5A). With data sets simulated from *rbcL* and analysed by NJ, the data sets containing 141 taxa proved much more difficult to reconstruct than the matrices with 357 and 567 taxa, with only 50% of the nodes correctly inferred versus 63 and 64% respectively (Fig. 2.4A). This discrepancy between data sets with different number of taxa for the first codon position of *rbcL* is not present in the MP analyses (Fig. 2.5A). Finally, analyses of simulated data sets based on the third codon positions of *atpB* and *rbcL* correctly recovered ca. 85% of the nodes with 10,000 bp using either NJ (Fig. 2.4C) or MP (Fig. 2.5C). However, the 141 taxa data sets for *rbcL* were again much more problematic, but this time both MP and NJ were affected and could only recover between 70 and 75% of the nodes (Figs. 2.4C and 2.5C).

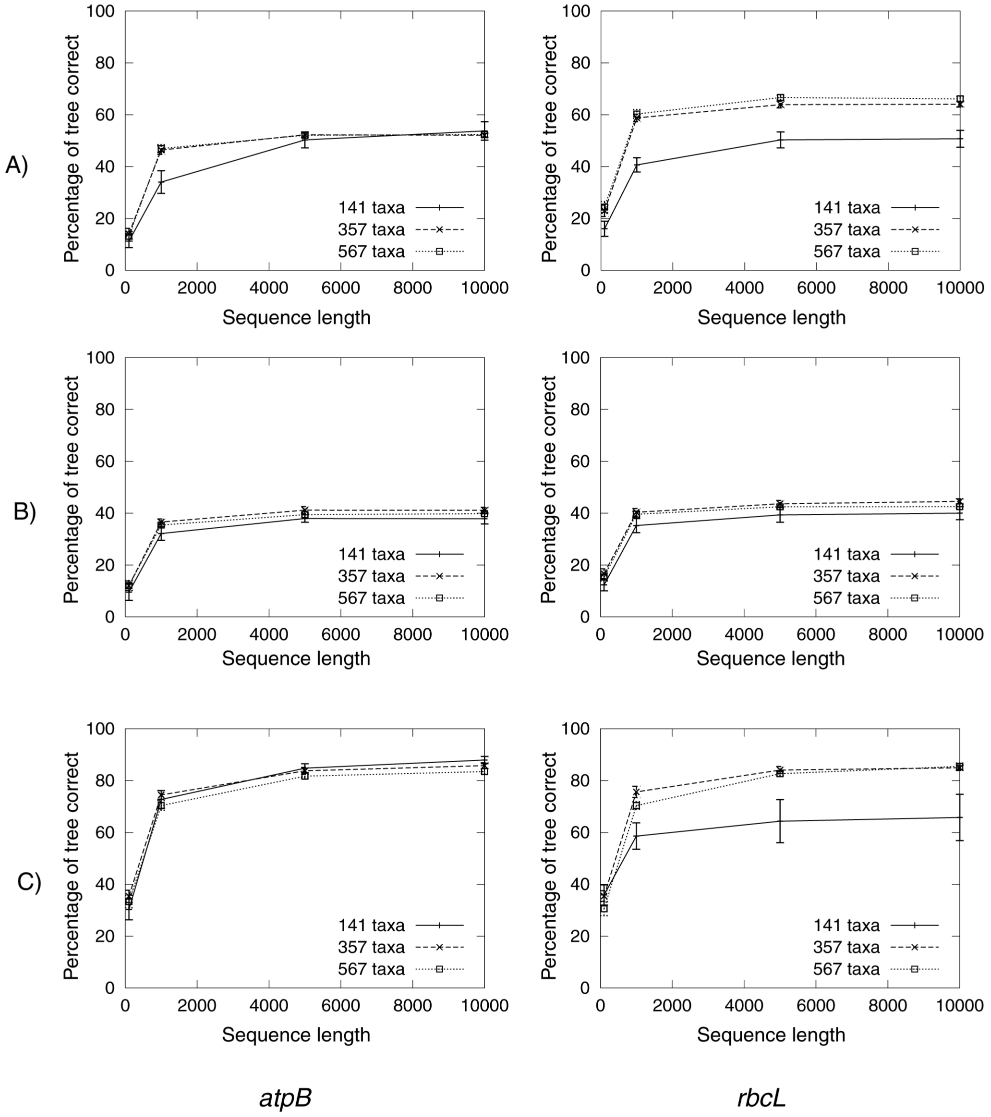


Figure 2.4 - Efficiency of NJ (HKY85+Γ distance) for estimating the different model trees from the simulated data sets based on the three codon positions of atpB *and* rbcL. *Vertical bars represents standard deviation around the mean. A) first codon position, B) second codon position and C) third codon position.*

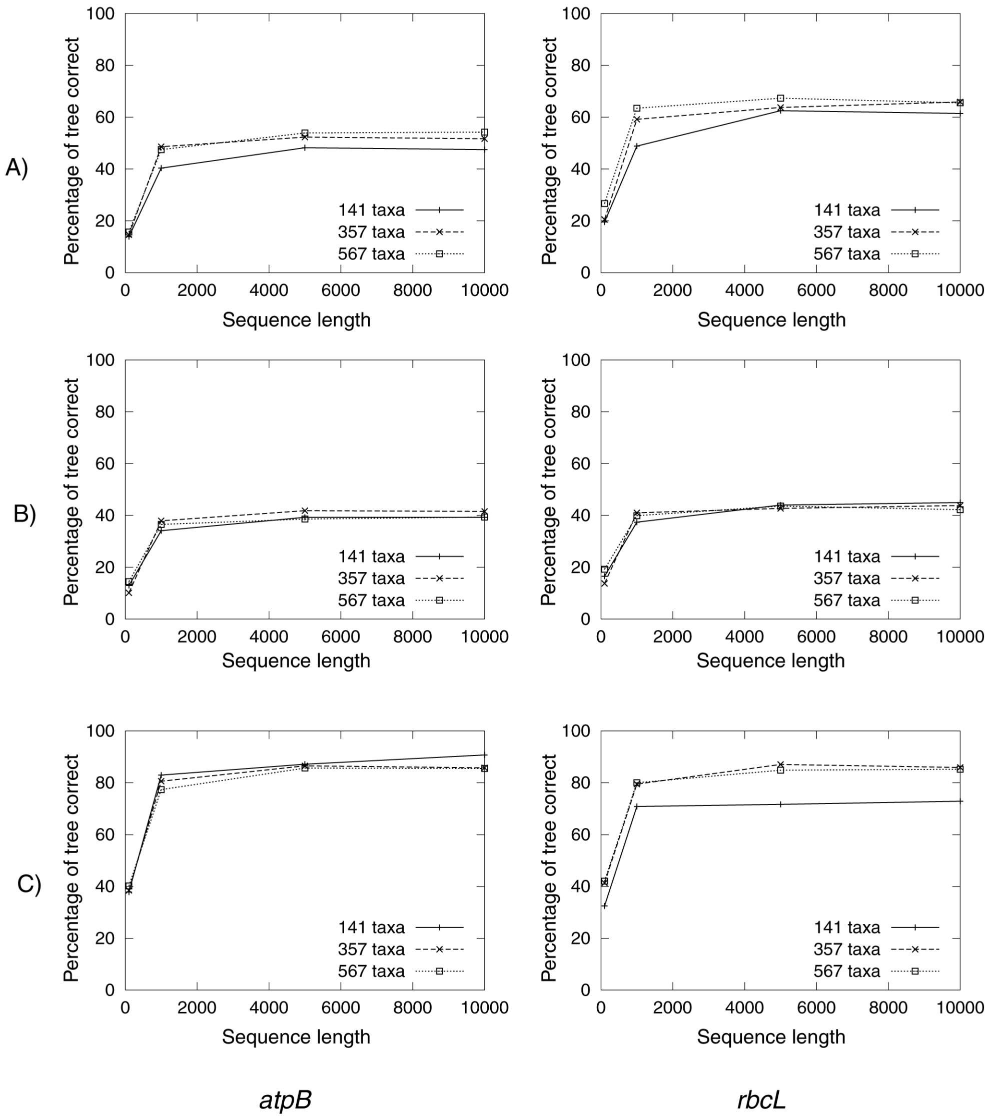


Figure 2.5 - Efficiency of MP for estimating the different model trees from the simulated data sets based on the three codon positions of atpB *and* rbcL. *A) first codon position, B) second codon position and C) third codon position.*

2.3.5 A 13,000 taxa tree

Results of the simulations based on the 13,000 taxa tree and analysed with MP and NJ methods are shown in Figure 2.6. NJ analyses could not be performed for data matrices larger than 10,000 bp due to the limited time jobs could be run on the 32 node IBM NetFinity cluster (a maximum of 96 hours can be allocated to each job). Contrary to expectations, NJ analyses (both with K2P or HKY85+Γ distances) for such a large matrix required far more CPU time to complete than the MP searches. For example, a data matrix with 100 bp took 1h20:35 and 1h15:30 of CPU time for MP and NJ respectively. With 10,000 bp, the time spent for the searches was 3h29:55 and 89h20:23 for MP and NJ respectively, while 30,000 bp were analysed in 5h19:41 by MP. It is not clear why NJ searches could not be completed on the NetInfinity cluster. The NJ algorithm compute a tree in a time that is function of the number of taxa present in the distance matrix (Swofford et al., 1996). Therefore, the same time should be spend with 10,000bp and 30,000bp, and the problem encountered here might come from the computation of the distance matrix. Further investigations are required to determine whether the problem is a computational problem (e.g. leak of memory in the software used) or an algorithmic one. At the same time, MP searches outperformed NJ with more than 1,000 bp (Fig. 2.6). With 10,000 bp, NJ correctly inferred 63% of the nodes from the 13,000 taxa tree, while MP correctly inferred 80% (Fig. 2.6). The percentage of tree correct continued to steadily increase with the addition of more characters to reach 82% with 30,000 bp with MP.

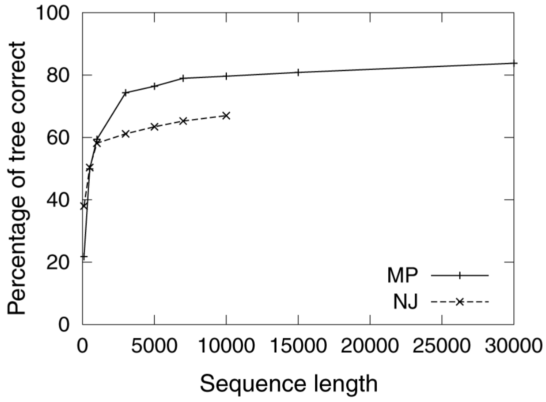


Figure 2.6 - Efficiency of MP and NJ (HKY85+Γ distance) for estimating the model tree from the simulated data sets containing 13,000 taxa. NJ performances dropped behind MP with sequences of 3,000 bp and more. The time required for completing the analyses was also much larger for NJ than for MP with large numbers of characters.

The number of characters that would be required to correctly infer 100% of the nodes from this 13,000 taxa tree was estimated from a logarithmic model fitted on the points obtained with MP (y = 8.461 * log(x); r^2 = 0.984; Fig. 2.6), leading to a total of 135,800 bp.

2.4 Discussion

Efficiency of two different phylogenetic methods to reconstruct four large angiosperm trees was assessed in this chapter by using computer simulations. Both MP and NJ performed surprisingly well under the conditions of the simulations despite the large number of possible trees existing for such large matrices (Figs. 2.1B and 2.2). The results obtained with these simulations mirrored those of Hillis (1996), suggesting that the 228 taxa tree for 18S rDNA used by Hillis (1996) is not a hit in the dark and several other large angiosperm trees can also be easily reconstructed by MP and NJ using different DNA sequences. The pattern of results was similar in each simulation performed, with a steep increase in the success of MP and NJ to reconstruct the model trees when the number of characters increased from 100 to 1,000 or 3,000 bp, followed by a plateau where the rise in percentages were minor with subsequent increases in sequence lengths (Figs. 2.1 to 2.6). This plateau occurred whether the trees, reconstructed from the simulated data sets, were close to the model trees or not and with each tree size investigated (i.e. 141, 228, 357, 567 and 13,000 taxa). It has been suggested that a ratio or value exists where the amount of characters is too low to allow tree reconstruction methods to discriminate between different tree topologies (e.g. Erdös et al., 1997; Kim, 1998). The simulations in this chapter seem to indicate that sequence lengths of less than 1,000 bp do not contain enough information to allow MP or NJ to successfully reconstruct the model trees used. Clearly different data sets of identical length can contain varying amounts of informative characters, but the point made here is a general one. The simulated sequences represent ‘perfect’ data sets that contain almost no constant or uninformative characters. With real data sets, the appropriate length of sequence will vary according to the DNA region and the organisms studied.

Simulations based on the 18S rDNA suggested that smaller trees containing 141 or 228 taxa were more efficiently reconstructed by the two methods used than larger trees with 357 and 567 taxa. The difference was, however small and more than 90% of the nodes were correctly inferred in the two larger trees (Fig. 2.1B). This difference could suggest that by adding more taxa, the problem of selecting the optimum solution from the tree space is getting harder. This is however not always the case, and the smaller 141 taxa tree proved to be much more difficult to reconstruct for NJ and MP than the two larger ones with simulations based on *atpB* and *rbcL* (Fig. 2.2). Adding more taxa to a phylogenetic analysis can be seen as a strategy to reduce the impact of long-branch attraction, thus avoiding the pitfalls of the Felsenstein zone (e.g. Purvis and Quick, 1997). Such a strategy has been used in chapter 5 of this thesis to demonstrate that inference of grass phylogenies has been impaired by long-branch attraction problems for some gene regions. However, in order to be of any use, the additional taxa have to be judiciously selected in order to intercept long branches (Graybeal, 1998). The four angiosperm trees used in this chapter represent an increasing sample of the flowering plant families, but it is unclear whether the conclusions reached by Graybeal (1998) can be applied to the phylogenetic problems investigated here. The estimated shape of the distribution of branch lengths for each model tree used in the simulations (Table 2.2) did not show large changes between the different sizes of trees for each data partition. Although larger trees tended to have a larger shape parameter, which could indicate more similar branch lengths, this remained the case whether model trees were based on 18S rDNA or *atpB* and *rbcL* (Table 2.2).

An important aspect not taken into account when estimating the distribution of branch lengths is whether adjacent branches are of similar lengths or not. The ratios of parent/daughter branch lengths calculated on each model tree suggested that the model tree for *atpB* with 141 taxa had a higher proportion of small internal branches giving birth to a daughter branch of at least four times its length than any other model tree considered (56.83%, 48.04% and 49.71% for 141, 357 and 567 taxa trees respectively; Table 2.4). The difference between the 141 model trees and the two other trees was more accentuated when the internal branches with daughter branches of at least twice their length were considered (75.73%, 64.11% and 63.99% for 141, 357 and 567 taxa trees respectively; Table 2.4). However, the number of internal branches grows when the number of taxa increases, and so should the probability of inconsistently estimating an internal branch (Kim, 1996). This seems to be in contradiction with the results found here for *atpB*. The problem of heterogeneous branch length was first investigated by Felsenstein (1978a) in an unrooted four-taxa tree containing two long and three short branches. The two nodes of this unrooted tree had two daughter branches of different length (i.e. one long and one short), while the length of the internal branch connecting these two nodes was similar to the two short terminal branches. This heterogeneity induced MP to consistently group the taxa at the tip of the two long branches together when the difference between long and short branches was sufficiently large and/or the rate of substitution was large enough (Felsenstein, 1978a). Kim (1996) expanded the case to trees with large numbers of taxa. He also considered quartets, but the terminal taxa of Felsenstein (1978a) were replaced by subtrees containing any number of terminal taxa. In such a configuration, he showed that the length of the five branches defining the quartet was not of primary importance, but that the total length of the subtree is the determinant factor. The measurement used here does not take this effect into account.

Hillis' (1996) explanation for the success of the different methods on 228 taxa (Fig. 2.1B) was that homoplasy in the data was distributed over the many branches of the tree, thus making unlikely any covarying patterns of homoplasy between any two taxa. The simulations performed with the 18S rDNA using the published trees for 141 and 567 taxa and the pruned tree for 320 taxa (Fig. 2.1A) seem to corroborate Hillis' (1996) explanation. Indeed, the performances of MP and NJ were much lower on these 'suboptimal' trees than on the trees specifically reconstructed with 18S rDNA (Fig. 2.1). The largest difference was found with 320 taxa, with the topology that was obtained by pruning 37 taxa from the published 357 matrix of Savolainen et al. (2000). Therefore, the topology used to create the simulated data sets had an impact on the success of the different methods to reconstruct the model trees, with 'suboptimal' topologies for the respective DNA sequence being used to estimate the parameters of the model of evolution proving more difficult than topologies obtained directly from the respective DNA sequence. However, the difference between these topologies was not reflected in the distribution of branch lengths (i.e. the number of long and short branch lengths; Table 2.2), but by an increase in the smaller ratios of each parent/daughter branch lengths for 'suboptimal' topologies (Table 2.4) and so by the distribution of these long and short branches along the tree.

Third codon positions performed much better than the other two positions (Figs. 2.4 and 2.5), and this difference was detected with both *rbcL* and *atpB*. The evolutionary rate between the three codon positions is very different, and the mean number of observed substitutions varied by almost ten- and five-fold between second/third and first/third codon positions (Table 2.3). The fast evolving third positions have often been down-weighted or excluded because of saturation of phylogenetic information contained when evolutionary rates are high (Swofford et al., 1996). However, the simulations performed here indicated that data sets based on third codon positions were easier to reconstruct for MP and NJ. One possible problem with the first two codon positions was that estimated branch lengths were too small to produce enough variability to reconstruct accurately the phylogenetic tree. Purvis and Quicke (1997) investigated the effect of increasing rate of evolution on the 18S rDNA data sets of 228 taxa and showed that the performances of MP started to decrease only when rates rose by a factor of 20. At this level, the sequences showed virtually no obvious homology (Purvis and Quicke, 1997). However, the rate of evolution *per se* did not seem to be the cause of the success of MP and NJ in reconstructing the model trees. Indeed, the evolutionary rate of first codon positions is similar to 18S rDNA, but less than 60% of the nodes were correctly inferred with simulated data sets based on this codon position (Figs. 2.4A and 2.5A). This is in comparison to a figure of more than 90% with 18S rDNA (Fig. 2.1B). At the same time, the rate of evolution of *atpB* and *rbcL* is almost four times that of 18S rDNA, but similar results were obtained with these different DNA sequences (Fig. 2.1 and 2.2). It seems however that the distribution of changes along the branches of the tree had a high impact on the performances obtained with each codon position. The estimates of the branch length distributions showed a pronounced L-shape distribution for second codon positions, which lessened for the first codon positions and became identical to the 18S rDNA for the third codon positions (Table 2.2). At the same time, second codon positions had a high proportion of parent/daughter ratios smaller than 0.25 (between 68.36 and 73.18; Table 2.4). First codon position had between 55.31% and 66.66% of ratios smaller than 0.25, while third codon positions had values similar to the complete coding genes (Table 2.4).

Almost all families of angiosperm were sampled in the 567 taxa matrix, and a logical next step would be to obtain sequences for all genera of angiosperm, which would represent 13,000 taxa. Likewise, building the tree of life would require the sampling of large numbers of taxa if inconsistencies caused by incomplete taxon sampling are to be avoided (Gauthier et al., 1988; Donoghue et al., 1989; Farris et al., 1996). Reconstructing such large phylogenetic trees could possibly require many more characters than those needed for 567 taxa (Soltis et al., 1998), and the size of such data matrices will imply that calculations get more demanding and time consuming. A large phylogenetic tree for the land plants containing 2,538 taxa (Källersjö et al., 1998) has been built using parsimony jackknifing (Farris et al., 1996), demonstrating the feasibility of such attempts. However, it is always possible to build a tree, but it is more difficult to know if that tree is close or not to the ‘true’ tree. The simulations performed with 13,000 taxa in this chapter are reassuring (Fig. 2.6). Although the branch lengths assigned to the 13,000 taxa topology in the simulations are probably not representative of real biological branch lengths (except for the initial distribution used), MP correctly inferred 82% of the nodes with a sequence length of 15,000 bp (Fig. 2.6). The heuristic search option used was crude, and it is possible that more thorough searches would increase the percentages obtained. However, this would come at the expense of time used to perform the searches (i.e. TBR swapping was tried without swapping on multiple trees but the search could not be finished within 96 hours). The extrapolation obtained from the percentages of tree correctly inferred by MP suggested that more than 100,000 bp could be required for such a large tree to be correctly inferred. However, it is likely that better and faster algorithms than the one used here could find a correct tree with less characters (Ronquist, 1998; Quicke et al., 2001).

Finally, NJ was selected in order to be able to perform several replicates under each of the simulated conditions, and the HKY85+Γ distance used matched the model of evolution selected to simulate the data sets thus taking into account the potential multiple hits that would impair MP searches (Felsenstein, 1978a; Swofford et al., 1996). However, MP was found to give slightly better percentages than NJ in most circumstances. One potential problem affecting NJ was demonstrated by Strimmer and von Haeseler (1996) who showed that accuracy and more importantly, for these simulations, average similarity between a model tree and the NJ tree decreases with an increase in number of taxa. Although the simulations done on matrices up to 567 taxa were performed much more quickly with NJ than with the heuristic search used for MP, this was not the case with the 13,000 taxa tree. It was noteworthy, in addition to the much longer computational time required by NJ, that computer memory was the restriction for NJ and searches could not be performed on our desktop computers such as an eMac G4 with 800 Mb of RAM. The IBM 32 node cluster was the only solution to perform these searches. This was the case whether the HKY85+Γ or the K2P distance were selected. On the other hand, MP searches were still possible on an eMac G4.

2.5 Conclusion

Increasing the sequence length up to 1,000 or 3,000 bp had a great impact on the success of the two method of phylogenetic reconstruction used to infer the model tree, and the addition of taxa from 141 to 567 did not result in major decrease in the performances of MP and NJ. Although the success found with the 228 taxa example of Hillis (1996) could be repeated with different angiosperm trees and using different DNA sequences using MP and NJ, the distribution of the expected changes along the tree was found to have an impact on the performances of each method. However, when the conditions are right, simulations of data sets containing 13,000 taxa showed that simple heuristic searches can give surprisingly good results and correctly infer a high proportion of nodes from data sets greater than 10,000 bp. However, one aspect that the simulations did not show, and that the researcher should consider, is what to expect from the analysis. A large number of taxa could be grouped perfectly well with only a few characters, and the expected resolution, also very low, might be satisfactory for some purposes. At the same time, even if 90% of the nodes were correctly inferred, it is important to consider which nodes are correct and whether they represent important clades, for which the phylogenetic tree was inferred.

With real data sets, it is not possible to know whether the tree(s) obtained are close or not to the true underlying phylogeny for the taxa at hand. However, techniques, such as the bootstrap or the jackknife, have been proposed to estimate how much confidence one can put in the trees inferred. These techniques are computationally intensive, especially when dealing with large number of taxa, and the next chapter investigates the effects of various heuristic searches on the level of support obtained for a large angiosperm DNA matrix.

CHAPTER 3. ASSESSING INTERNAL SUPPORT WITH LARGE PHYLOGENETIC DNA MATRICES FOR THE ANGIOSPERMS

3.1 Introduction

The production of a phylogenetic hypothesis involves typically two steps. Firstly, a method of phylogenetic inference, such as ME, ML, MP is used to produce phylogenetic trees. Unlike the computer simulations performed in chapter 2, it is impossible to know in advance the true underlying evolutionary history for the taxa under investigation. It is, however, very important with real data sets to be able to quantify how much confidence we can have in the hypothesis represented by the topologies obtained. Secondly, measures of internal support are calculated to discriminate between clades with clear phylogenetic signal and those needing further work (i.e., more data, perhaps involving a search for factors creating incongruence). Various methods have been proposed to assess internal support, but resampling techniques like the bootstrap and the jackknife are by far the most commonly used; for mathematical details, see Efron (1979) or Efron and Tibshirani (1993); for their application to phylogenetic analyses, see Felsenstein (1985a) and Penny and Hendy (1985). Other methods such as relaxation of parsimony or ‘decay’ indices (Bremer, 1988) or double decay analysis (Wilkinson et al., 2000) have also been proposed. Bayesian analysis has been recently developed and is seen as a promising method for estimating confidence on tree topologies, especially because of its relative rapidity for a model-based method (i.e. in comparison to bootstrap searches under ML). Posterior probabilities have the advantage of clearly quantifying uncertainties in the tree topology directly from the MCMC sampling algorithm. Efron et al. (1996) gave a mathematical description of bootstrap estimation, and showed that bootstrap percentages can be considered identical to posterior probabilities obtained by assuming a flat Bayesian prior on the data-space. A two-level bootstrap algorithm is required to obtain percentages that can be interpreted in a similar way to classical confidence intervals (Efron et al., 1996).

3.1.1 Searching the tree space

When dealing with analyses that include hundreds of taxa, the search through the tree space to find the optimal tree(s) is a computationally intensive procedure that can take years of CPU time (e.g. Rice et al., 1997). The assessment of internal support generally requires several hundreds of replicates, each replicate having to perform a complete heuristic search. Therefore, most researchers have assessed internal support by using a fast algorithm to bootstrap or jackknife their data, which for MP and ML methods may involve no branch-swapping. Chapter 2 has shown that even simple heuristic searches on a large number of taxa can correctly recover more than 95% of the nodes present in a model tree when the number of available characters increases. But for a data matrix of a given length, as more rigorous swapping algorithms are employed, further improvements on the initial tree obtained are common (i.e. they are shorter), and thus the expectation is that the tree search, as well as the bootstrap/jackknife percentages might be strongly dependent on the heuristic tree search method used. Furthermore, it has always been difficult to evaluate what minimum percentage is required to consider a clade ‘well supported’.

3.1.2 Aims

In this chapter, a large data matrix has been used to compare the bootstrap and jackknife percentages of various heuristic searches using several of the swapping algorithms available. By examining differences in internal support at each node between these different methods, the aim is to provide some guidelines when assessments of support for large matrices are being carried out.

A previous study (Mort et al., 2000) covered similar topics, but they took a different approach to that used here. First, a much larger matrix of DNA sequences was used (2925 characters *x* 357 taxa in our study vs. 1494/1060 characters *x* 24/84 taxa respectively in Mort et al., 2000) to assess how well these methods performed under such conditions (performance of these methods has been well characterized with the small data sets of Mort et al., 2000). More importantly, a particular interest was to determine if there was a clear point at which unclear patterns became clear, such that some form of natural breakpoint could be detected to separate ambiguous results, characterized by wide variance in percentages obtained with different methods, from clear patterns in which consistent support was obtained regardless of method used. The thinking here was that perhaps there was some point at which enough data points were present to create more consistent support percentages and that such a point could be used as the cut-off for a clade being ‘well supported’.

3.2 Material and Methods

3.2.1 Data and phylogenetic analyses

The data set of Savolainen *et al.* (2000) that included the *atpB* and *rbcL* plastid genes (2925 characters x 357 plant taxa) was analysed. Taxon nomenclature follows recommendations from a recent re-classification of the families and orders of angiosperms (Angiosperm Phylogeny Group, APG, 1998). One hundred bootstrap and jackknife (50% character deletion) resamplings were performed using PAUP*4b. A simple taxon addition was used and each of the following with five trees held at each replicate: no swapping, NNI, subtree pruning and regrafting (SPR), and TBR swapping algorithms. To evaluate the effects of keeping limited number of trees, another bootstrap search was also performed with the NNI swapping algorithm but keeping 100 instead of five trees per replicate. To evaluate the effect of the number of replicates, 1000 bootstrap replicates were performed using NNI swapping and keeping five trees per replicate.

3.2.2 Statistical analyses

Regression analyses were performed to evaluate the effect of the number of replicates and the number of trees kept per replicate on the bootstrap analyses. The support values obtained by the different heuristic searches were normalized using the arcsin transformation (i.e. arcsin of the square-root of each percentage; Sokal and Rohlf, 1995). Pearson correlation coefficients were then calculated for all these methods (Sokal and Rohlf, 1995). Nodes were then ranked according to the bootstrap percentages obtained with the TBR swapping algorithm and grouped into 12 percentage intervals (i.e. <50, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85-89, 90-94, 95-99 and 100%). Two-way analyses of variance followed by multiple comparisons tests (Tukey test; Sokal and Rohlf, 1995) were performed on each interval to investigate if the methods gave more similar results when the support for the node was higher and if a cut-off point could be found. Pooled means and confidence intervals for each interval were also calculated for the set of swapping options for which difference in mean did not deviate significantly from zero based on the multiple comparison test (Sokal and Rohlf, 1995). All the statistical analyses were performed using the software R version 1.4.0[5].

[5] http://www.r-project.org

3.3 Results

Percentages obtained with 100 replicates of bootstrap and NNI were plotted (Fig. 3.1) against percentages obtained with the same swapping algorithm with 1000 replicates. A high correlation exists between the two analyses (slope =1.002, r^2 = 0.947, $p < 0.001$; Fig. 3.1) showing that support obtained with only 100 replicates gave an excellent approximation, particularly for higher bootstrap percentages to a search with 100 replicates.

Percentages obtained when saving five versus 100 trees per replicate with the number of replicate held constant were also highly correlated (slope = 0.987, r^2 = 0.942, $p < 0.001$; Fig. 3.2). Depending on the swapping algorithms used, the percentages differed among the eight methods from no difference at all (e.g., *Citrus/Poncirus*, Haloragaceae/*Penthorum*; Appendix 3.1) up to ca. 20-30% (e.g., Brassicales/Malvales, Gentianales/Lamiales; Appendix 3.1). For the three swapping algorithms, the more rigorous algorithms gave consistently higher bootstrap/jackknife percentages. TBR provided on average the highest bootstrap/jackknife percentages; those with SPR slightly lower but still close to TBR. NNI gave the lowest percentages. However, this was not always the case, for example, nodes separating *Cichorium/Menyanthes* or *Pistacia/Schinus* received a slightly higher percentage with NNI (Appendix 3.1). When no swapping was used, bootstrap/jackknife percentages were on average 10% lower than the worst percentages with any type of swapping. In some cases, the absence of swapping led to extremely low percentages for the largest clades (e.g., 4% and 9% versus 99% and 88%, respectively, for rosids and eurosid I; Appendix 3.1).

Figure 3.1 - Bootstrap percentages above 50 with 100 replicates using NNI swapping plotted against bootstrap percentages for the same clades with 1000 replicates (see text; slope = 1.002, r^2 = 0.947, $p < 0.001$). A high degree of correlation exists between the different bootstrap analyses, particularly for the higher bootstrap values.

(See figure next page)

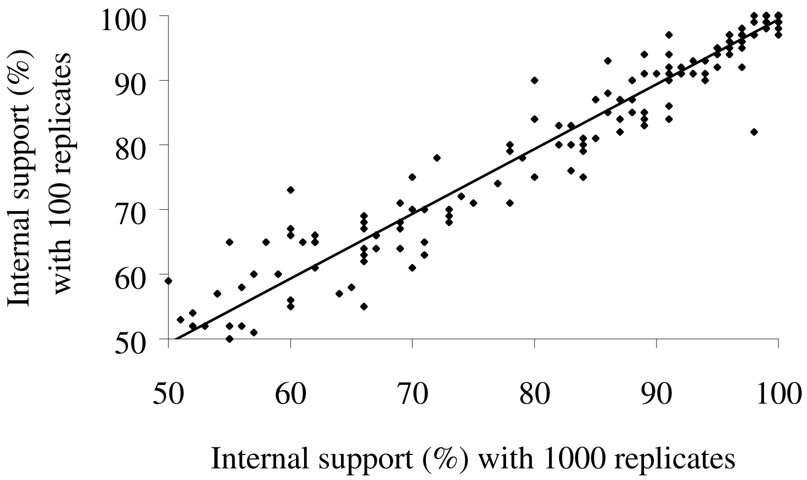


Figure 3.1

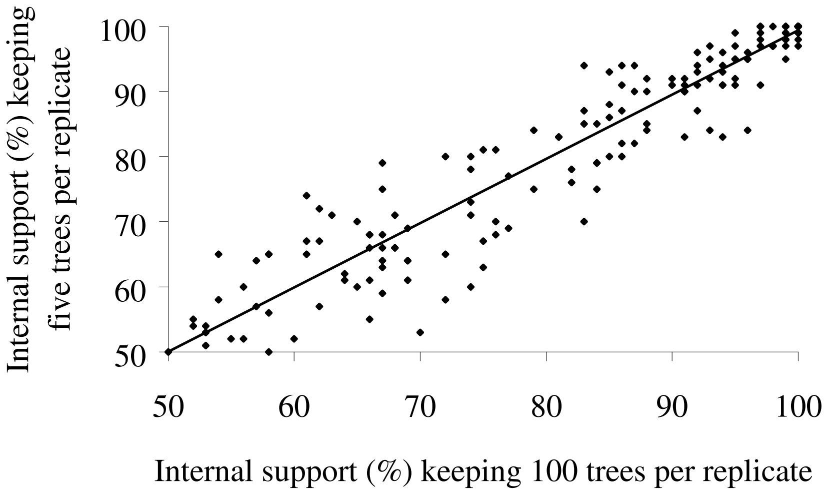


Figure 3.2 - Bootstrap percentages above 50 with 100 replicates using NNI swapping and keeping up to 100 trees at each replicate plotted against bootstrap percentages for the same clades keeping five trees at each replicates (see text; slope = 0.987, r^2 = 0.942, p < 0.001). A high correlation exists between the two heuristic searches used to obtain bootstrap percentages particularly for the higher bootstrap values.

There was no observable trend for the comparisons between bootstrap and jackknife resampling techniques because neither can be seen to be producing consistently higher support percentages than the other (Appendix 3.1). To quantify the relationships between these different bootstrap/jackknife percentages, Table 3.1 provides Pearson correlation coefficients between the methods. All comparisons resulted in high correlations (correlation coefficients > 0.84). The higher correlations were found between the bootstrap and jackknife within each swapping option (Table 3.1), whereas no swapping had consistently lower correlation coefficients when compared to any of the other algorithms. The next step was to look into intervals of percentages after dividing the nodes into 12 groups. The two-way analyses of variance indicated that the bootstrap and jackknife (50% character deletion) resampling techniques gave similar information (Table 3.2). However, the various swapping algorithms gave significantly different percentages in all but two percentage intervals (Table 3.2). Percentages below 50 and between 70 to 74 had comparable levels of internal support for all swapping algorithms used (Table 3.2). No significant interaction was found between the type of resampling techniques and swapping algorithms used at a significance level of 5% (Table 3.2). As shown by Tukey' s multiple comparisons tests (Table 3.3), most of the differences between swapping algorithms found in the two-way analyses of variance were due to the no-swapping option. No swapping, with either the bootstrap or jackknife, typically produced significantly different percentages when compared to more rigorous swapping algorithms like TBR or SPR; this was not the case when no swapping was compared to NNI (Table 3.3). None of the three swapping algorithms (i.e. TBR, SPR or NNI) could be distinguished from the others by Tukey' s test (Table 3.3).

The mean obtained for each group of support with each method is presented in Figure 3.3. The pooled mean and confidence intervals obtained after the Tukey' s tests are also shown (Fig. 3.3). TBR did not always give the highest percentages, and none of the three swapping algorithms consistently produced the highest percentages. However, an increase in similarity of results for the various methods was observed when higher percentages (above 90%) are considered (Fig. 3.3), but these are not significantly different from the other categories.

Table 3.1 - Pearson correlation coefficients between the eight different heuristic search options for 100 bootstrap or jackknife replicates keeping 5 trees at each replicate. All comparisons resulted in high correlations, with no swapping algorithm having the lowest correlation coefficients.

		Bootstrap				Jackknife		
		NNI	SPR	TBR	No swapping	NNI	SPR	TBR
Bootstrap	SPR	0.94	1.00	---	---	---	---	---
	TBR	0.93	0.95	1.00	---	---	---	---
	No swapping	0.92	0.88	0.87	1.00	---	---	---
Jackknife	NNI	0.96	0.91	0.92	0.92	1.00	---	---
	SPR	0.89	0.89	0.91	0.85	0.90	1.00	---
	TBR	0.93	0.95	0.94	0.89	0.92	0.91	1.00
	No swapping	0.91	0.86	0.86	0.96	0.92	0.86	0.87

Table 3.2 - Table of F values from two-way analysis of variance for the eight heuristic search options for each subgroup of percentages. No significant difference was found between the two resampling techniques, while swapping algorithms gave significantly different percentages in all but two percentage intervals.

Group of percentages	Swapping		Resampling		Interaction	
	F	*p*-value	F	*p*-value	F	*p*-value
< 50	2.27	0.11	2.92	0.10	2.50	0.08
50-54	8.25	<0.001	0.02	0.89	0.35	0.78
55-59	7.88	<0.001	0.02	0.90	0.43	0.73
60-64	13.18	<0.001	0.14	0.71	0.08	0.97
65-69	6.32	<0.001	0.50	0.48	0.74	0.53
70-74	1.69	0.17	0.14	0.71	0.35	0.79
75-79	8.22	<0.001	0.05	0.82	0.22	0.88
80-84	19.74	<0.001	0.07	0.80	0.10	0.96
85-89	15.10	<0.001	0.62	0.43	0.12	0.94
90-94	18.87	<0.001	1.66	0.20	0.45	0.72
95-99	31.46	<0.001	0.07	0.79	0.13	0.94
100	21.53	<0.001	0.06	0.81	2.03	0.40

Table 3.3 - Table of Q values ('studentized range') resulting from Tukey' s test of multiple comparisons based on the eight different heuristic search options investigated. Values indicating a significant difference (P < 0.05) are marked by an asterisk (*) and bold face. Bootstrap: tbrb = TBR; sprb = SPR; nnib = NNI; nswb = No swapping. Jackknife: tbrj = TBR; sprj = SPR; nnij = NNI; nswj = No swapping.

Comparisons	Percentages									
	50-54	55-59	60-64	65-69	75-79	80-84	85-89	90-94	95-99	100
tbrb/nswj	4.52	3.53	**5.75***	3.31	3.27	**4.92***	4.27	**5.66***	3.44	2.69
tbrb/nswb	4.22	2.98	**5.65***	2.88	4.04	**4.84***	4.16	4.11	3.56	2.87
tbrb/nnij	2.98	0.90	2.50	0.55	2.52	1.88	1.71	1.99	1.12	1.26
tbrb/nnib	1.66	0.22	1.81	1.75	2.27	2.10	1.19	1.93	1.44	1.20
tbrb/sprj	1.12	0.62	0.24	0.51	0.14	0.86	0.30	0.35	0.45	0.69
tbrb/sprb	0.19	0.93	0.33	0.87	0.41	0.95	0.50	0.13	0.58	0.56
tbrb/tbrj	0.24	1.20	0.38	0.40	0.40	0.48	0.07	0.57	0.03	0.72
tbrj/nswj	**4.76***	**4.73***	**5.37***	2.91	2.87	**4.44***	4.34	**5.09***	3.42	1.97
tbrj/nswb	4.46	4.18	**5.27***	2.48	3.64	**4.36***	4.23	3.54	3.56	2.14
tbrj/nnij	3.22	2.10	2.11	0.15	2.12	1.40	1.78	1.42	1.12	0.53
tbrj/nnib	1.90	1.42	1.42	1.35	1.87	1.62	1.26	1.36	1.41	0.47
tbrj/sprj	0.88	0.58	0.14	0.90	0.26	0.38	0.36	0.22	0.48	0.04
tbrj/sprb	0.04	0.27	0.05	0.47	0.81	0.47	0.43	0.44	0.60	0.16
sprb/nswj	**4.72***	**4.46***	**5.42***	2.45	3.68	3.97	**4.77***	**5.53***	4.01	2.13
sprb/nswb	4.42	3.91	**5.31***	2.01	**4.45***	3.89	**4.66***	3.98	4.14	2.30
sprb/nnij	3.17	1.83	2.16	0.32	2.93	0.93	2.22	1.86	1.70	0.70
sprb/nnib	1.86	1.15	1.47	0.88	2.68	1.15	1.69	1.81	2.01	0.63
sprb/sprj	0.92	0.31	0.09	1.37	0.55	0.08	0.80	0.22	0.12	0.12
sprj/nswj	**5.64***	4.15	**5.51***	3.82	3.13	4.06	3.97	**5.31***	3.90	2.01
sprj/nswb	**5.34***	3.60	**5.41***	3.39	3.90	3.99	3.87	3.76	4.02	2.18
sprj/nnij	4.10	1.51	2.26	1.05	2.38	1.02	1.42	1.86	1.58	0.57
sprj/nnib	2.78	0.84	1.57	2.26	2.13	1.24	0.89	1.81	1.89	0.51
nnib/nswj	2.86	3.31	3.95	1.56	1.00	2.82	3.08	3.72	2.01	1.50
nnib/nswb	2.56	2.76	3.84	1.13	1.77	2.75	2.97	2.17	2.13	1.67
nnib/nnij	1.32	0.68	0.69	1.20	0.25	0.22	0.52	0.06	0.31	0.06
nnij/nswj	1.54	2.63	3.26	2.76	0.75	3.04	2.56	3.66	2.32	1.44
nnij/nswb	1.24	2.08	3.15	2.33	1.53	2.97	2.45	2.11	2.44	1.61
nswb/nswj	0.30	0.55	0.10	0.43	0.77	0.07	0.10	1.55	0.12	0.17

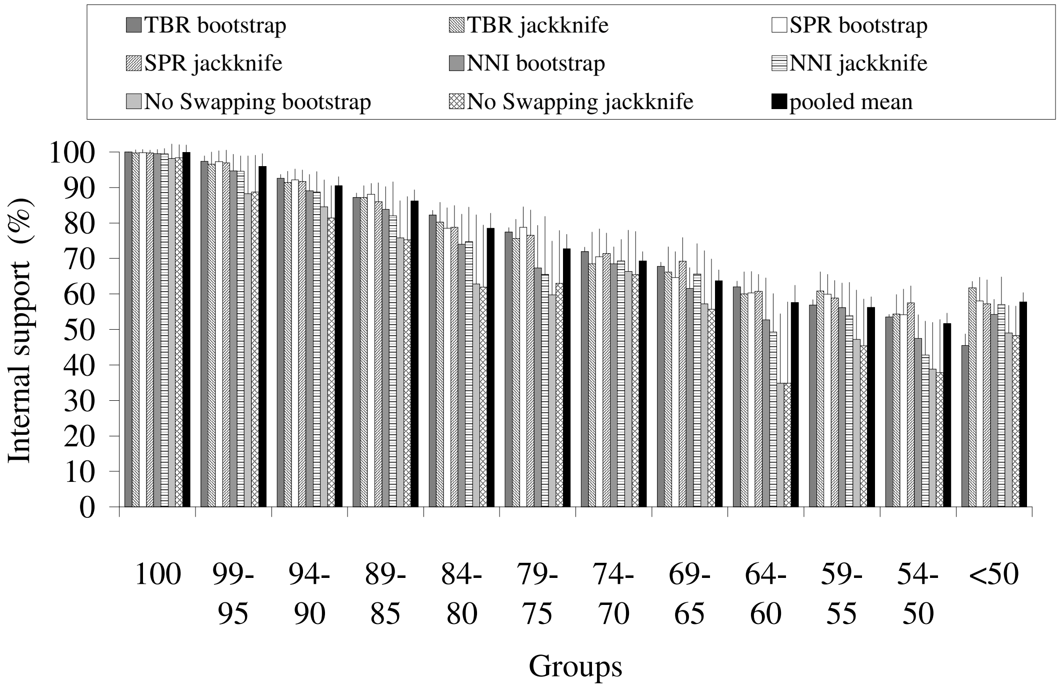


Figure 3.3 - Mean percentages and standard error of (ordered from left to right) TBR bootstrap, TBR jackknife, SPR bootstrap, SPR jackknife, NNI bootstrap, NNI jackknife, No swapping bootstrap, no swapping jackknife, and pooled mean and 95% confidence intervals for the 12 groups of node support. The pooled mean was calculated for the methods that could not be distinguished with Tukey' s test (p < 0.05; Table 3).

3.4 Discussion

These comparisons show that the results of bootstrap and jackknife analyses with various swapping algorithms are highly correlated when applied to a large matrix, particularly at higher support percentages (above 90%). It is, however, not possible to define a clear threshold at which all methods gave consistent percentages. Felsenstein (1985a) first applied the bootstrap to phylogeny reconstruction and stated that because confidence limits on a statistic are frequently constructed by the percentile method, all clades occurring in 95% or more of the bootstrap estimates should be searched (the empirical upper and lower 2.5% points of the distribution of bootstrap estimates of the statistics). Hillis and Bull (1993) argued that the bootstrap confidence percentages were biased and too conservative as an assessment of internal support, but Efron et al. (1996) showed that Felsenstein' s method provided a reasonable first approximation for actual confidence levels of the observed clades (see also Berry and Gascuel, 1996). However, because more rigorous swapping options will search more thoroughly the tree space for any given data set, the 95% interval *per se* cannot be a clear cut off because this interval varies according to swapping algorithms. Because NNI, SPR, and TBR swapping algorithms are correlated and cannot be distinguished statistically in the data set analysed here (Tables 3.1 and 3.3), some percentages slightly lower than 95% using NNI would receive at least 95% with TBR (e.g. Rubiaceae, Malpighiales, Fabales, Ranunculales, etc.; Appendix 3.1).

Although no clear cut off was observed at a certain level of support at which all methods gave similar percentages, above 95% all methods including no swapping agreed. A similar conclusion was reached by DeBry and Olmstead (2000) with simulations. To our surprise, this was also the case for the below 50%, 65-69% and 70-74% intervals. Percentages below 50 were represented by only four nodes in the data, and the similarity found between the methods used could be due to that artefact. The last interval is somewhat perturbing because it is the only interval for which analysis of variance found no significant differences between the eight methods, although the number of observations was similar in all groups. We expected to find such *p*-values only with higher percentages, but this was not the case (Table 3.2). Only two methods gave different support percentages, and these were the bootstrap and jackknife methods with no swapping. These differences can be extremely large (see Fig. 3.3 and Appendix 3.1), which casts doubt on the reliability of these fast algorithms. Mort et al. (2000) found that searches without swapping gave similar results to searches performed using NNI, which conclusion could reflect the effect of large data sets on reconstruction methods (Sanderson and Wojciechowski, 2000).

The number of replicates in bootstrap or jackknife analyses has been suggested to influence internal support, but in these results all percentages (Fig. 3.1) were highly correlated, so one hundred replicates provided satisfactory results. This is particularly true for high percentages, whereas low percentages are found to be less reliable. This finding contradicts the recommendations of Hedges (1992), who argued that 1,000 replicates were required for accurate estimation of bootstrap percentages. As showed by Mort et al. (2000), increasing the number of replicates resulted in a decrease of the standard error around the mean nodal support. However, Felsenstein (1985a) stated that even with a small number of bootstrap replicates a general estimate can be constructed for which parts of a phylogenetic estimate are well supported, a conclusion in agreement with the results presented here.

Some might consider the jackknife superior to the bootstrap. At the time Felsenstein was addressing this question, jackknife estimates consisted of dropping one character at a time and then estimating the phylogeny: the resulting trees varied far less than did those in bootstrap estimates (Felsenstein, 1985a). Felsenstein (1985a) proposed that one way to make the jackknife vary as much as the bootstrap would be to drop half the characters chosen at random. Farris et al. (1996) and Swofford (2000) provided such a tool, especially the parsimony jackknifer (Farris et al., 1996), which has been used recently with complex phylogenetic trees (Källersjö et al., 1998 for plants; Lipscomb et al., 1998 for eukaryotes). The comparison of the jackknife (50% character deletion) and the bootstrap (Table 3.1) clearly shows that both methods gave similar results (Pearson correlation coefficients > 0.85), and thus one can be substituted for the other. Variations in results produced with the bootstrap and jackknife are far less than that caused by use of the various swapping algorithms and are not significant (Fig. 3.3; Table 3.2).

Finally, these results have particular significance if dealing with large DNA matrices. With several hundred taxa, all heuristic searches will take a long time, as exemplified by the study by Rice et al. (1997), who spent nearly one year of CPU time re-analysing a data set composed of 500 DNA sequences (Chase et al., 1993). Even though a new, faster algorithm has appeared for large data sets (Ronquist, 1998), performing searches on each of the reconstructed fictional (bootstrap) or partitioned (jackknife) sets of data remains time-consuming. However, bootstrapping with NNI took roughly 15 times faster than SPR and 30 times faster than TBR on the angiosperm matrix used in this chapter. Because these results have shown that percentages derived from these algorithms are not statistically different with the large angiosperm matrix, it is sufficient to use the faster algorithms as long as the target percentages have been corrected according to what we can reasonably expect from a particular swapping algorithm. However, using no swapping during the heuristic search, even though it is much faster and thus more tractable than NNI, can give significantly different percentages, and this method cannot be recommended. It is also crucial that authors provide explicit information regarding the bootstrap protocol applied, particularly in oral presentations when such information is typically omitted. For example, it would be misleading to say that monophyly of the euasterids (a large clade) did not receive support >50% if it was also not stated that the authors used the fast jackknife option of PAUP*4b, which uses no swapping (14% jackknife value; Appendix 3.1). This same euasterid clade received more than 60% when the data were bootstrapped using TBR (Appendix 3.1). Here, the results demonstrated that no matter which swapping algorithm was used, bootstrap and jackknife percentages were all correlated. What is of prior importance is to define clearly what range of percentages are expected to provide evidence of support, instead of focusing on some absolute percentages (such as >50%, or as 95%) regardless of the particular behaviour of the swapping algorithms used. The narrowing of percentages over 90% obtained with the various methods could be used as the range of percentages to indicate ‘well supported’, although at the same time it should be noted that this phenomenon occurred in other, much lower ranges as well and was not significantly more narrow than the others (even if 65-69% and 70-74% were excluded from the calculations). In the end, it must be admitted that a degree of arbitrariness is involved in choosing a range of percentages to represent good support.

3.5 Conclusion

The resampling techniques used to estimate clade reliability on phylogenetic trees are computationally intensive and assessing the effects of faster versus thorough heuristic search options on the support estimates obtained is important in order to reduce the time spend on these procedures. At the same time, such comparisons can also give insights on how to deal with published phylogenetic trees where support estimates have been obtained by different procedures. This could help comparing support estimates between studies, but it can also provide a way to use this information with a meta-analysis approach. Bootstrap or jackknife support values can be incorporated into methods used to produce supertrees and this type of application is discussed in the next chapter.

CHAPTER 4. BUILDING SUPERTREES: AN EMPIRICAL ASSESSMENT USING THE GRASS FAMILY

4.1 Introduction

Comprehensive and well-resolved phylogenetic trees, containing when possible estimates of divergence dates, underpin comparative biology and allow powerful tests of a wide range of hypotheses to be made (Felsenstein, 1985b; Harvey and Pagel, 1991; Pagel, 1999). Phylogenetic trees representing a large sample of taxa are also preferred to those based on limited taxa number for classification purposes and for studying character evolution.

4.1.1 Building large phylogenetic trees

Two different approaches can be used to obtain comprehensive phylogenetic trees. The first uses characters gathered from the widest possible range of taxa directly in an analysis to produce a ‘big tree’. In this approach, phylogenetic analyses of true biological characters, either molecular or phenotypic (e.g. morphological), are performed, and the meaning of the evolutionary hypotheses underlining these characters can be interpreted. This approach has been investigated in chapter 2 and 3, and can be considered as the classical approach used to infer phylogenetic relationships. The second is the meta-analysis approach (Arnqvist and Wooster, 1995) used in supertree building methods. The underlying idea of these methods is to combine topologies (or source trees) resulting from multiple phylogenetic studies (Sanderson et al., 1998), rather than their respective biological data sets, to produce a supertree. Matrix elements derived directly from these published topologies, and for which no real phylogenetic interpretation can be obtained, represent the characters that are used to build the supertree or composite phylogenetic tree (Bininda-Emonds and Bryant, 1998). While consensus techniques also work on topologies in order to produce a summarized phylogenetic tree, supertree reconstruction has the advantage of not requiring identical terminal taxa sets. Only overlapping sets are needed allowing the method to produce more comprehensive phylogenies than the original ones.

A recent theoretical study (Hillis, 1996), as well as the simulations performed in chapter 2, suggested that large phylogenetic trees can be easier to analyse than previously thought and empirical analyses have demonstrated that large combined multi-gene analyses can correctly infer large trees (Soltis et al., 1999; Savolainen et al., 2000; Soltis et al., 2000). Because ‘big trees’ may not be as easy to construct as suggested by these studies (chapter 2 of this thesis; Kim, 1996; Purvis and Quick, 1997) and because combinable data are not always available, the use of composite phylogenies to study evolutionary patterns is one of the few choices left to the investigators. However, choosing between the various supertree methods is not a straightforward task.

4.1.2 Supertree reconstructions

Several basic algorithms have been proposed for supertree reconstruction, and Steel et al. (2000) gave an account of desirable properties required for a supertree method. The Strict Supertree Reconstruction method (Constantinescu and Sankoff, 1995) or the OneTree algorithm (Bryant and Steel, 1995) attempts to directly assemble the topologies of the source trees into a supertree, and requires compatible trees (i.e. trees without conflicting nodes) as input. Incompatible trees cannot be incorporated in these methods, and no solution to this problem has been found (Steel et al., 2000). This limitation precludes its use with real data sets, where incompatible nodes, reflecting either real evolutionary divergence between data sets (hard incongruence) or simply random error due to limited sample size (soft incongruence), are often the rule between different trees. The MinCut algorithm (Semple and Steel, 2000; Page, submitted) is a promising modification of the OneTree algorithm allowing the building of supertrees from incompatible source trees. More importantly, it computes a supertree in polynomial time (Bryant and Steel, 1995; Semple and Steel, 2000; Page, submitted). However the algorithm can be sensitive to the size of the source trees, and can fail to include information that is not contradicted in the set of input trees, therefore more work is needed to improve the algorithm.

In contrast, Matrix Representation with Parsimony (MRP - Baum, 1992; Ragan, 1992) can be used whether or not the source trees are compatible. A drawback of MRP method is that it can not find a supertree in polynomial time (i.e. it is NP-complete), and therefore can be trapped in the same problems as the classical phylogenetic methods when large number of taxa are involved (see chapter 2). MRP uses additive binary coding (Farris et al., 1970) to represent the hierarchical structure of trees as a series of matrix elements (Baum and Ragan, 1993). Every node on each source tree is represented by a binary matrix element and MP analysis of the matrix is used to retrieve the tree(s) that represent(s) the hierarchical information in the source trees. If multiple most-parsimonious composite trees are obtained, a strict or semi-strict consensus can be used to generate a consensus composite tree (Baum, 1992; Ragan, 1992). A similar approach as MRP can be obtained by computing the patristic distance between each taxon from each source tree to create a distance matrix instead of a binary matrix (Lapointe and Cucumel, 1997; M. J. Sanderson, pers. comm.), but this approach is not considered here.

4.1.3 Coding procedures

Several coding procedures have been proposed for the MRP method. Baum (1992) and Ragan (1992) independently first proposed that terminal taxa delimited by each node should be coded as '1' in the binary matrix, and all other taxa as '0'. Missing taxa from individual source trees are then coded as missing values (typically '?') for the matrix elements representing these trees. Purvis (1995) argued that the elements derived from source trees lack independence, and hence add redundant information to the matrix. He proposed removing this apparent redundancy by allocating the value '0' only to taxa within the immediate sister group to the particular clade under consideration, and by assigning missing values to the other taxa of the source tree (i.e. coding them as '?' instead of '0'). Subsequently, Ronquist (1996) suggested that the bias would not be associated with redundant information as stated by Purvis (1995), but with the relative sizes of the source trees. He argued that the difference in the amount of information contributed by each source tree could be removed by inversely weighting each tree according to its number of nodes. However, he favoured other weighting schemes based on the support for nodes, which, he argued, would also compensate for any size bias.

Bininda-Emonds and Bryant (1998) and Bininda-Emonds and Sanderson (2001) discussed some of the properties of MRP and investigated modifications to the method. Little is known about the merits of such modifications or how they perform with real data. Bininda-Emonds and Bryant (1998) also discussed the issue that supertrees obtained from MRP are not always congruent with those based on an approach using combined data, and suggested that different weighting schemes might help MRP to approximate the combined result more closely. Furthermore, matrix elements represent membership (character state coded as '1') or lack of membership (character state coded as '0') of a particular taxon relative to a clade. Allowing reversals in the MP analyses can therefore produce clades in the composite tree that are supported by a lack of membership in some components of the source trees. Bininda-Emonds and Bryant (1998) advocate using irreversible character states in a MP analysis to overcome this shortcoming.

4.1.4 Aims

In this chapter, the results of an empirical study aimed at assessing the relative merits of the supertree approach using the grass family as a case study are reported. The data set was a combined molecular and morphological data set from the Grass Phylogeny Working Group (GPWG, 2001) and it was used to compare the diverse modifications of the MRP methods that have been proposed, to investigate the effect of irreversible characters on supertree reconstructions, and to evaluate the differences between the supertrees we obtained and an approach using combined data. Accurate and meaningful comparisons between the combined analysis and the supertree approach were made by breaking down the GPWG combined data set into its eight character partitions and then rebuilding a phylogenetic tree using the MRP modifications. The same MRP modifications were then used to produce different large supertrees including 401 genera from 55 published phylogenies and their strengths and weaknesses in relation to other evolutionary hypotheses concerning the grass family are discussed. These trees are among the largest ever produced for the grass family.

4.2 Material and Methods

4.2.1 Combined phylogenetic tree

The GPWG' s data set (GPWG, 2001), which combined eight different data sets for 61 species of grasses, representing molecular as well as morphological data, was reanalysed using MP and well-supported clades were determined using bootstrap percentages (Felsenstein, 1985a). One thousand bootstrap replicates were performed (using the TBR swapping algorithm with random addition of taxa and keeping 20 trees at each step) using PAUP*4b. The GPWG matrix was also divided into its eight data partitions representing three plastid gene sequences (*ndhF*, *rbcL* and *rpoC2*), three nuclear DNA regions (*gbss*, *phyB* genes, and 5.8S and internal transcribed spacer 2 of the rDNA, hereafter ITS), plastid restriction site variations as well as morphological data (Table 4.1). MP analyses were performed on each character set followed by 1,000 bootstrap replicates with the same heuristic search options as previously described.

Table 4.1 - Details of the eight character partitions of the GPWG data set.

Data type	Genome	Number of characters	Number of taxa
ndhF	plastid	2186	51
rbcL	plastid	1344	30
rpoC2	plastid	855	29
phyB	nuclear	1182	39
gbss	nuclear	773	14
ITS	nuclear	424	42
restriction sites	plastid	364	45
morphology	-	52	61

4.2.2 MRP reconstructions

Two different types of MRP analyses were performed for the grass family. For the two sets of supertree reconstructions, five binary matrices were built using the program SuperTree 0.85b[6] (Fig. 4.1). Firstly, the eight different bootstrap trees obtained from each data partition of the GPWG matrix were used as source trees for supertree reconstructions (referred hereafter as GPWG supertree(s)) with the same 61 terminal taxa as in the GPWG combined tree. Secondly, a total of 55 publications were considered for supertree reconstructions (referred hereafter as large supertrees) to produce matrices of 401 genera (out of a total of 635 for the whole family; Mabberley, 1993). The publications chosen do not represent an exhaustive sample of the published literature concerning the grass family, which was not the goal of our study. Given the homogeneity of characters and methods used in the reviewed publications (Appendix 4.1), no distinction was drawn between each of the source trees. Our decision also followed Bininda-Emonds et al. (1999) and Purvis (1995b), who found that differential weighting of the source trees according to data and/or tree search method had little impact on the composite phylogenetic tree.

[6] see Chapter 8. Technical notes for its description

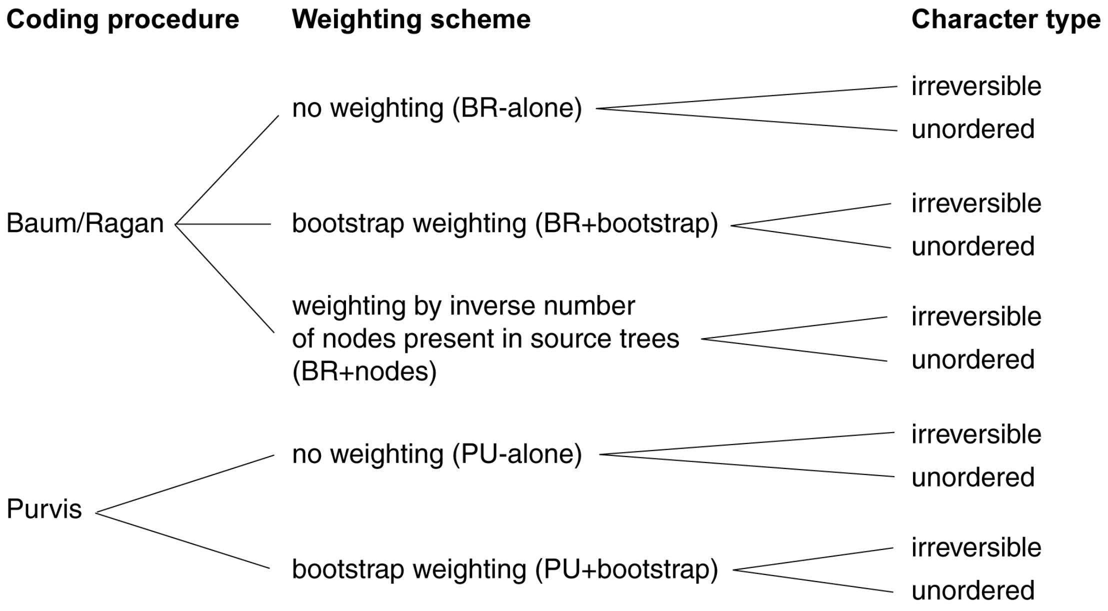


Figure 4.1 - Schematic representation of the different MRP coding procedures compared. Each coding procedure has been performed twice, once for the gpwg supertrees and once for the large supertrees.

Supertree reconstruction requires an overlap of taxa sampling between source trees. However, due to the large size of the family (about 10,000 species), few taxa were common between published studies. To overcome this problem, only generic names were considered in the large analyses. Species were used when evidence against the monophyly of the genus was demonstrated in the published study. Three taxonomic groups, pooids, panicoids and chloridoids, are consistently found as strong monophyletic groups and no evidence has ever contradicted this view (Clark et al., 1995; Duvall and Morton, 1996; Mason-Gamer et al., 1998; Gaut et al., 1999; Hilu et al., 1999; Hsiao et al., 1999; Mathews et al., 2000). In order to ease the heuristic searches, the large analyses of the 55 published phylogenetic trees were constrained by forcing each of these three groups to form three monophyletic clades, but allowed other taxa to be inserted within them (the ‘backbone’ option in PAUP*4b).

Because bootstrap percentages are missing in most studies before 1993, or values less than 50% are not specified for the majority of published trees, a transformed function of the bootstrap percentages to weight the characters into the coded matrices in both analyses was used. All percentages below 50%, or nodes with missing values, were given the weight of one. Bootstrap percentages above 50% (inclusive), were weighted using an exponential transformation (James S. Farris, pers. comm.):

$$e^{[\log(a)x(\frac{b}{100})]}$$

where *a* represents the weight assigned to 100% of node support, and *b* represents the bootstrap percentage. A value of 100 was assigned to *a* in the subsequent analyses performed. This transformation attempts to overcome the conservative bias found in bootstrap percentages. Indeed, Hillis and Bull (1993) showed, using simulations, that bootstrap values over 70% usually indicate, with greater than 95% probability, that the corresponding clade was real in their study. Corrections to bootstrap estimates have been proposed (Rodrigo, 1993; Zharkikh and Li, 1995; Efron et al., 1996) but these corrections cannot compensate for lack of information in large analyses (Sanderson and Wojciechowski, 2000). This is because as the number of taxa increases, the relative number of characters available declines and phylogenetic accuracy suffers. Our transformation attempts to linearize the bootstrap percentages, thereby allowing a gradual increase in the character weights in subsequent MP analyses.

For every supertree reconstruction, heuristic searches under MP were performed using both Baum/Ragan and Purvis coding schemes alone (hereafter BR-alone and PU-alone, respectively). Bootstrap percentages weighting schemes were used as described above on both coding schemes (hereafter BR+bootstrap or PU+bootstrap, respectively) and Baum/Ragan coding scheme weighted by the inverse of the number of nodes present in each source tree (i.e. each character derived from a source tree is inversely weighted by the number of nodes found in that source tree; hereafter BR+nodes). Analyses were performed with 1,000 replicates of random addition sequence using the NNI swapping algorithm with only 20 trees kept at each replicate. The stored trees were then swapped with the TBR swapping algorithm and a maximum of 1,000 trees were kept (the ‘maxtrees’ option in PAUP*4b. Two different MP analyses were performed in each of the cases described above: one considering characters as unordered (Fitch, 1971), and one considering the same characters as irreversible (Camin and Sokal, 1965), leading to a total of ten GPWG supertrees and ten large supertrees (Fig. 4.1). Equally most parsimonious solutions were summarized using semi-strict consensus.

4.2.3 Topological comparisons

Two incongruence indices, expressed as distances, were used to compare the different topologies obtained. The partition metric (PM; symmetric difference in PAUP*4b; Robinson and Foulds, 1981) and the agreement subtree metric (D1; Gordon, 1980) were calculated using PAUP*4b. For the comparisons between the combined analysis and the GPWG supertree reconstructions, Kishino-Hasegawa (KH-test, Kishino and Hasegawa, 1989) as well as Shimodaira-Hasegawa (SH-test, Shimodaira and Hasegawa, 1999) tests were performed. The GPWG data set consisted of DNA sequences, restriction site data and morphological characters (Table 4.1), and the two tests were performed under the MP criterion. Two sets of *p-value*s were thus calculated, one following the default options present in PAUP*4b (for the KH-test), and the other following the procedure described in Shimodaira and Hasegawa (1999) to ensure the validity of the test with *a posteriori* specified topologies. The latter test involved the creation of 500 bootstrapped replicates of the GPWG matrix followed by the optimisation of 1000 random trees as well as the trees under consideration on these bootstrapped matrices (see Shimodaira and Hasegawa, 1999 and Goldman et al., 2000 for details).

4.3 Results

4.3.1 GPWG combined tree

The GPWG combined tree (Fig. 4.2) is one of the most comprehensive phylogenetic hypotheses concerning the grass family. Several features can be emphasized and will serve as a reference for the supertree comparisons. Firstly, and following the system proposed by GPWG (2001), a large clade (PACCAD clade; Fig. 4.2) composed of six subfamilies - panicoids, arundinoids, chloridoids, centothecoids, aristidoids and danthonioids - formed a highly supported monophyletic group (100% bootstrap; Fig. 4.2). No bootstrap values above 52% supported any subfamilies within the PACCAD clade. Outside the PACCAD clade, the bambusoids, ehrhartoids, and the pooids formed another group called the BEP clade supported by a lower bootstrap value (87% bootstrap; Fig. 4.2). Finally, three clades - the anomochloids (*Anomochloa*, *Streptochaeta*), pharoids (*Pharus*), and puelioids (*Guaduella*, *Puelia*) - formed the basal taxa of the Poaceae (early-diverging lineages; Fig. 4.2).

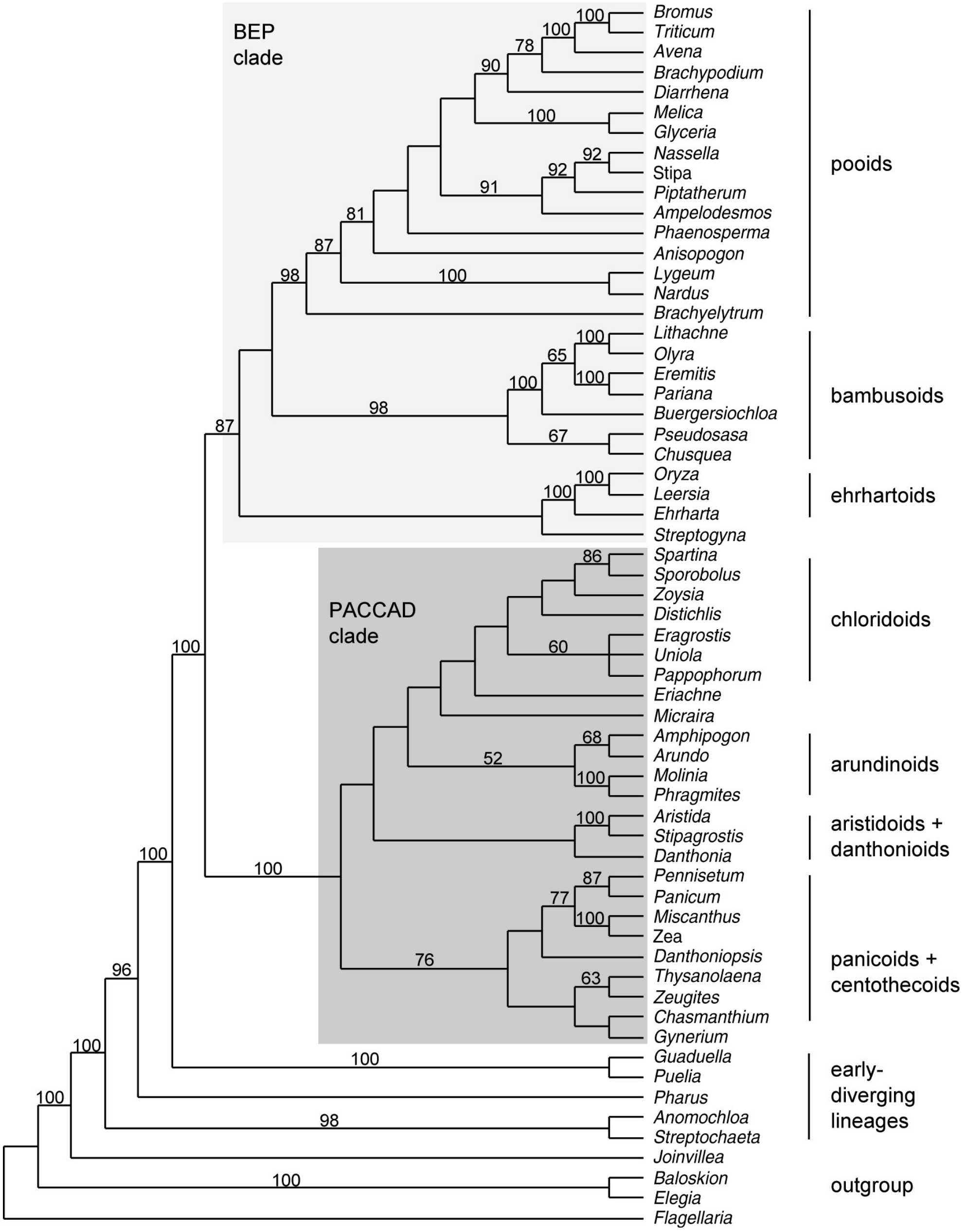


Figure 4.2 - The ‘big’ GPWG tree based on the analysis of eight different data sets (both molecular and morphological). Numbers above branches represent bootstrap percentages.

4.3.2 GPWG supertrees

The proportion of non-identical nodes between the ten MRP reconstructions taken by pairs was calculated using PM (Table 4.2). Analyses using unordered characters resulted in topologies with pairwise distances between 0.36 and 0.47 (mean distance: 0.41; Table 4.2A). Pairwise comparisons for each MRP modification using irreversible characters gave distances between 0.22 and 0.49 (mean distance: 0.36; Table 4.2B). Values for each unordered Baum/Ragan modification were closer to their irreversible counterparts than to any other methods (mean distance unordered BR/irreversible BR: 0.33; mean distance unordered BR/irreversible PU: 0.41; Table 4.2C). The same pattern was found for the unordered Purvis modifications (mean distance unordered PU/irreversible PU: 0.39, mean unordered PU/irreversible BR: 0.47; Table 4.2C). Three modifications gave very similar topologies with both types of characters. BR-alone gave the exact same topology, while the two BR+bootstrap and the two PU+bootstrap were very close to each other respectively (0.11 and 0.18; Table 4.2C).

For comparisons between the GPWG combined tree and the GPWG supertrees, D1 distances ranged from 0.32 for BR+bootstrap with irreversible characters to 0.59 for PU-alone with unordered characters, the other eight modifications having values from 0.42 to 0.52 (Fig. 4.3). When considering distances obtained with PM, BR+bootstrap with irreversible characters again gave the topology the most similar to the GPWG combined tree with a distance of 0.25. The worst modification was the BR+nodes with irreversible characters with a distance of 0.53, while the eight other modifications ranged from 0.41 to 0.50. The placement of the major subclades defined in the GPWG combined tree (Fig. 4.4) was further examined. Two modifications, BR+bootstrap and PU+bootstrap, both with irreversible characters, are the only modifications to obtain the same basal branching pattern as the combined GPWG tree (Fig. 4.4A). With all the other modifications using irreversible characters, two basal grasses, *Guaduella* and *Puelia*, are inserted inside the BEP clade. BR-alone placed the pooids as sister group to the PACCAD clade, while BR+nodes and PU-alone place a clade formed with the pooids and the ehrhartoids as sister group of the PACCAD clade (Fig. 4.4A). Using unordered characters, and except with the BR-alone modification, which was identical to the BR-alone with irreversible characters modification, all modifications had an odd basal branching pattern with the outgroup being inserted between two clades of the early-diverging lineages (Fig. 4.4B).

Table 4.2 - PM for the five MRP modifications, expressed as distances. A) Comparisons within modifications with unordered characters, B) within modifications with irreversible characters, and C) between modifications with unordered characters and modifications with irreversible characters.

A)		Unordered characters				
		BR-alone	BR+bootstrap	BR+nodes	PU-alone	PU+bootstrap
Unordered characters	BR-alone	-	.	.	.	.
	BR+bootstrap	0.46	-	.	.	.
	BR+nodes	0.38	0.39	-	.	.
	PU-alone	0.44	0.36	0.42	-	.
	PU+bootstrap	0.47	0.38	0.46	0.36	-
B)		Irreversible characters				
Irreversible characters	BR-alone	-	.	.	.	.
	BR+bootstrap	0.43	-	.	.	.
	BR+nodes	0.22	0.49	-	.	.
	PU-alone	0.29	0.43	0.22	-	.
	PU+bootstrap	0.41	0.38	0.44	0.38	-
C)		Irreversible characters				
Unordered characters	BR-alone	0.00	0.43	0.22	0.29	0.41
	BR+bootstrap	0.46	0.11	0.55	0.49	0.46
	BR+nodes	0.38	0.48	0.32	0.33	0.48
	PU-alone	0.44	0.46	0.50	0.42	0.47
	PU+bootstrap	0.47	0.42	0.52	0.46	0.18

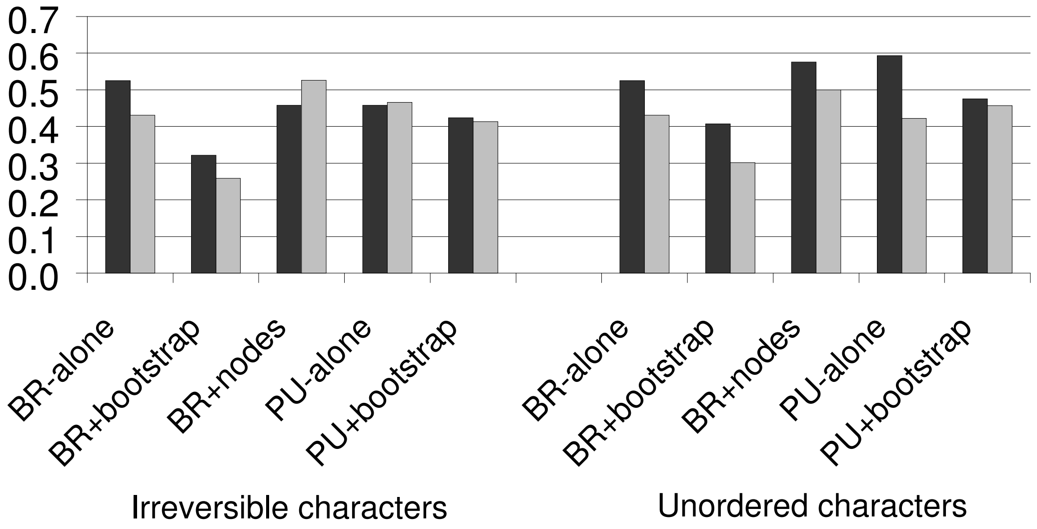


Figure 4.3 - D1 (black) and PM (gray), expressed as distances to the GPWG tree, for each of the GPWG supertrees build from separate phylogenies based on the eight character partitions of the GPWG combined data set as source trees with characters considered as irreversible (left) and unordered (right).

The two different implementations of the KH-test were used to estimate the validity of the ten GPWG supertrees as possible alternative hypothesis to the GPWG combined tree (Table 4.3). Based on the characters of the GPWG combined data set, the classical KH-test rejected all the GPWG supertrees (*p*<0.001) as being suitable alternative phylogenetic trees. However, the modified KH-test found only the PU-alone with unordered characters as being significantly different (*p*<0.05) from the GPWG combined tree (Table 4.3). All the other GPWG supertrees could not be rejected, BR+bootstrap with both characters types being once again the modification giving the closest topology to the GPWG combined topology (*p*=0.59 and *p*=0.62 for unordered and irreversible characters respectively).

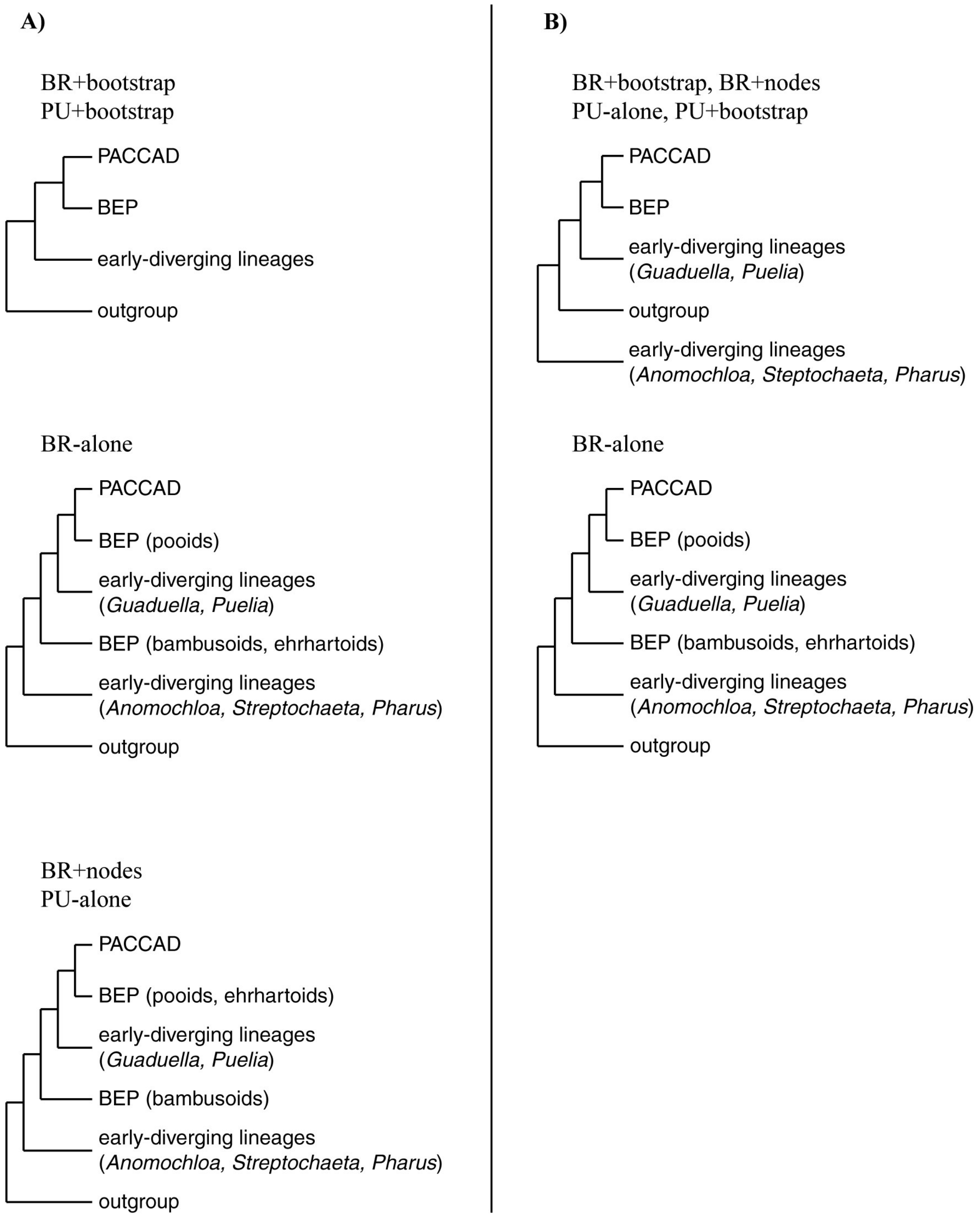


Figure 4.4 - Summary of the subfamilial relationships in the GPWG supertrees obtained with the five MRP modifications with column A) characters considered as irreversible and column B) characters considered as unordered.

*Table 4.3 - KH test based on the difference in length between the GPWG combined tree and the ten GPWG supertrees obtained after applying the different MRP modifications. The p-values were calculated using the default parameter of PAUP*4b (classic KH-test), and Goldman et al. (2000) modification to ensure the validity of the test with a posteriori specified topologies (modified KH test). Asterisk (*) indicates a significant p-value (p<0.05)*

			p-value	
		Number of steps	Classic KH test	Modified KH test
	Combined GPWG	9054	-	-
Unordered characters	BR-alone	9291	<0.001*	0.17
	BR+bootstrap	9116	<0.001*	0.59
	BR+nodes	9203	<0.001*	0.39
	PU-alone	9411	<0.001*	0.04*
	PU+bootstrap	9211	<0.001*	0.37
Irreversible characters	BR-alone	9291	<0.001*	0.17
	BR+bootstrap	9105	0.0003*	0.62
	BR+nodes	9281	<0.001*	0.19
	PU-alone	9311	<0.001*	0.14
	PU+bootstrap	9191	<0.001*	0.42

Finally, Figure 4.5 shows the supertree obtained with BR+bootstrap using irreversible characters. The discrepancies between this supertree and the GPWG combined tree were small and concerned mainly single taxa that were always positioned close to the clade they belonged to in the GPWG tree (Fig. 4.5). The main differences concerned the placement of the ehrhartoids and bambusoids that exchanged their position as sister group of the pooids in the supertree in comparison to the GPWG tree, but the relationships within these two groups are similar in both analyses. The placement of the three subfamilies arundinoids, aristidoids and danthonioids.

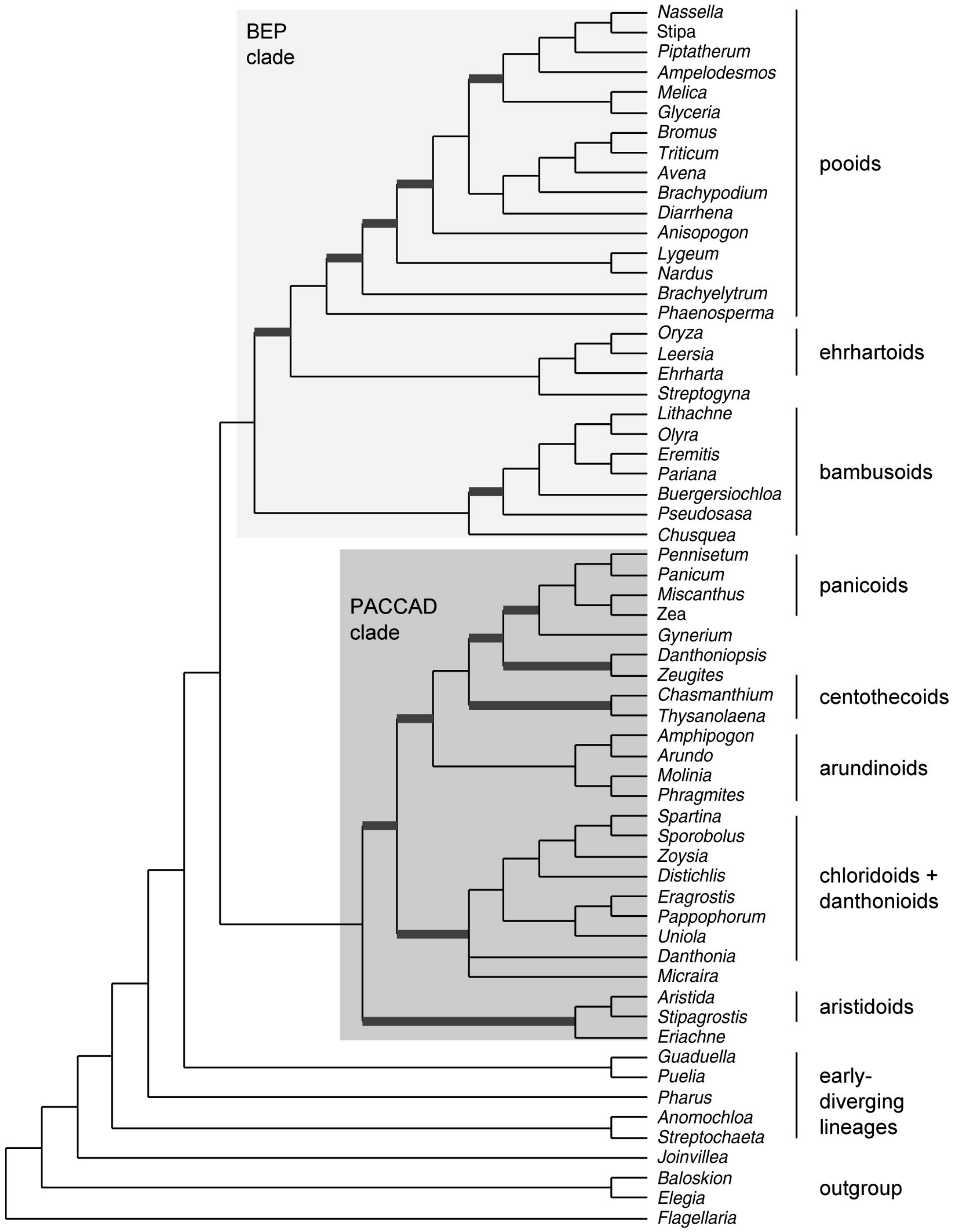


Figure 4.5 - Grass supertree obtained with Baum and Ragan coding scheme with bootstrap support weighting using separate phylogenies built from the eight character partitions of the GPWG combined data set as source trees. Bold lines indicate incongruent branches between the supertree and the GPWG combined tree.

4.3.3 Large supertree

The 55 publications were analysed in the same way as the eight character partitions from the GPWG combined data set. The results of BR+nodes are not shown for clarity, and because this MRP modification resulted in topologies incompatible with the placement of subfamilies as suggested by the GPWG combined tree. Using irreversible characters (Fig. 4.6A), BR-alone, BR+bootstrap and PU+bootstrap obtained the same basal branching pattern, but the placement of subfamilies inside the PACCAD clade was slightly different in each case[7]. PU-alone was the only modification to place the three subfamilies of the BEP clade together as a sister group to the PACCAD clade, so following the GPWG supertrees and the GPWG combined tree. Using unordered characters (Fig. 4.6B), BR-alone, BR+bootstrap and PU-alone obtained the same basal branching pattern, with the BEP clade being paraphyletic. PU+bootstrap gave an odd combination with the early-diverging lineages embedded inside the bambusoids at the base of the tree (Fig. 4.6B).

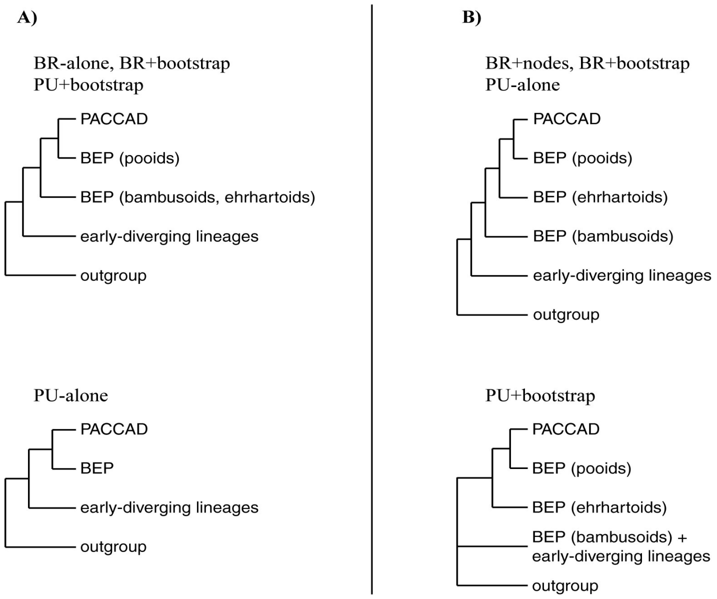


Figure 4.6 - Summary of the subfamilial relationships in the large supertrees based on 55 published source trees with column A) characters considered as irreversible and column B) characters considered as unordered.

[7] The five supertrees can be found on the CD-ROM attached

4.4 Discussion

The information given by the two incongruence indices helped pinpoint the differences present in the topologies obtained from the ten GPWG supertrees built using the eight character partitions from the combined data set. Analyses using unordered characters produced supertrees that were less similar, as measured by D1, to the combined tree than analyses using irreversible characters. The only exception is for BR-alone, which obtained the same topology with both character types. However, unordered characters produced slightly better topologies than irreversible characters using the PM incongruence index. D1 is defined as the number of taxa needed to be removed from both trees in order to get an identical subtree, while PM is the number of taxa bipartitions found between the two trees compared (Johnson and Soltis, 1998). Higher D1 than PM values suggest that the topological differences did not involve single taxa (in which case, D1 would be smaller than PM), but rather that a large set of taxa is placed differently in the supertree and the GPWG combined tree. Similar low values of D1 and PM can be obtained if only a few taxa are misplaced (leading to low D1 distance), with their location being close in both trees, which would not result in many wrong bipartitions (leading to low PM distance).

The D1 distances were much higher with unordered than irreversible characters, and these values were much closer to the PM ones when using irreversible characters. Following this logic, it is possible to emphasize that using unordered characters resulted in topologies with more differences from the GPWG combined tree, in the placement of larger subclades, than irreversible characters. The type of characters (i.e. unordered or irreversible) used to reconstruct the supertrees did not greatly influence the placement of the terminal taxa in the GPWG supertrees. Therefore, topologies obtained by considering matrix elements as irreversible characters were closer to the combined phylogenetic tree represented by the GPWG combined analysis. This effect is clearly visible in the placement of the major subclades defined in the GPWG combined tree (Fig. 4.5). The discrepancies found between the GPWG combined analysis and the GPWG supertree (Fig. 4.5) concerned the placement of clades that are problematic and where the taxon sampling remains low. The MRP approach can be related to a taxonomic congruence approach (Bininda-Emonds et al., 1999), and by such is likely to treat the phylogenetic signals found in the data partitions in a different way as a combined analysis. When combining data sets, the phylogenetic signal found in a data partition can be added to the signal from another partition to reinforce the support for a particular topology. In contrast, with the MRP method, the binary coding procedure remove any conflicting signal from a data partition by only representing the best topology, and the presence of two incongruent topologies in the source trees can weaken the signal for the possible supertrees. Unordered characters produced topologies where the early-diverging lineages were split into two groups, one including *Puelia* and *Guaduella* and one including *Anomochloa*, *Streptochaeta*, and *Pharus*. This is an extremely odd and unrealistic pattern, which does not correspond to any published phylogenetic trees concerning the grass phylogeny. Bininda-Emonds and Bryant (1998) found only a minor impact of irreversible compared to unordered characters in their analysis, but the topologies created using irreversible characters were closer to an approach using combined data than topologies created using unordered characters. The resolution in the large supertrees was extremely dependent on the type of characters used to perform the analysis, which was not the case with the GPWG supertrees where both types of characters gave similar resolution (Table 4.2). Using irreversible characters produced only one or two large supertrees depending on the modification used, while the 'maxtrees' option in PAUP*4b was always reached when using unordered characters. Of course, using irreversible characters should put more constraints on the MP analysis by preventing the reversion from state '1' to '0' thus reducing the number of most parsimonious trees, but it is not clear why such a drastic difference appeared only between the large supertrees.

The Purvis coding scheme has been proposed as an improvement to the Baum/Ragan procedure in order to reduce the dependency and redundancy between elements coming from the same topology (Purvis, 1995; Ronquist, 1996; Bininda-Emonds and Bryant, 1998). However, our results suggested that the Purvis coding scheme does not have a large impact on the MRP reconstructions and that the Baum/Ragan method works as well or even better. The comparisons between the approach using combined data and BR-alone and PU-alone gave similar values for both PM and D1 (Fig. 4.3). Comparisons between MRP modifications also indicated a close relationship between the two modifications especially with irreversible characters (Table 4.2). This conclusion is less well supported with unordered characters, because BR-alone is the only modification to give a very different basal topology to any other modification (Table 4.2 and Fig. 4.3). Ronquist' s (1996)proposition to weight each character in the binary matrix by the inverse of the number of nodes present in the corresponding source tree does not perform as well as the other modifications. It has the highest partition metric distance and the second highest agreement subtree distance (Fig. 4.3). Although the two alternatives to Baum/Ragan method are based on logical and plain arguments (i.e. non-independence and redundancy of matrix elements, and impact of larger trees), their effects on MRP reconstruction are not obvious and do not result in topologies closer to our combined reference. PU-alone with unordered characters was even rejected as a suitable alternative hypothesis to the GPWG combined tree with the modified KH-test and obtained the second smallest *p-value* when used with irreversible characters (Table 4.2). However, conclusions from the large supertree analysis differed, because PU-alone was the only modification to produce the same basal branching pattern as the GPWG combined tree (Fig. 4.6B), while BR-alone, with the pooids sister group to the PACCAD clade, corresponded to an alternative hypothesis for the grass family supported by some data from nuclear (*Adh*; Gaut et al., 1999) and plastid (*rbcL*; Duvall and Morton, 1996) genomes. Because no reference phylogenetic trees containing a similar number of taxa as our large supertrees are available, it is difficult to assess their topologies and to determine if PU-alone with irreversible characters really gave a more accurate large grass phylogeny than the other methods. The placement of the pooids is possibly not the best criterion to judge the methods as conflicts exist concerning its evolutionary position within the grass family inferred with different DNA sequence data sets.

Weighting by node support, as suggested by Ronquist (1996), improved the fit between the additive binary matrices and the GPWG combined tree. The beneficial impact of the bootstrap weighting scheme is evident from the increased *p-value*s obtained with the modified KH-test (Table 4.2). This trend is visible for the Baum/Ragan coding scheme, but is less obvious with the Purvis coding scheme (Fig. 4.3). Bootstrap weighting was proposed (Ronquist, 1996) as an alternative to the Purvis coding scheme in order to reduce, on the one hand, the bigger impact of large source trees over smaller ones, but also to improve the effect of well-supported nodes in the MRP analysis. It is not clear how bootstrap weighting could reduce the impact of larger trees if the smaller trees do not have much higher support values than larger trees. However, weighting of characters by bootstrap support within the Purvis coding scheme appears to be a redundant procedure unable to greatly improve the MRP reconstruction from the important information present in bootstrap support values (Fig. 4.3). In the large supertree analysis, PU+bootstrap with unordered characters even produced unrealistic topologies with the early-diverging lineages embedded in the bambusoids (Fig. 4.6B). The Purvis coding schemes removes important restrictive information from the matrix (Bininda-Emonds and Bryant, 1998), and it is possible that weighting by bootstrap support would randomly assign high values to character in the matrix where this restrictive information have been removed, preventing the weighting schemes to be as effective as with the Baum and Ragan coding schemes. Weighting the matrix elements by node support also poses problems. Not all bootstrap analyses can be considered as identical, and the number of replicates and/or the type of searches done will influence the support found. Moreover, when node support is not provided for a source tree, weights cannot be assigned and information from this tree is down-weighted in the MRP analysis. This has a great impact when supertree reconstruction is done with older publications, which in general do not have node support and are hence down-weighted in the MRP analyses.

An important aspect affecting the large supertrees that is not resolved in our comparisons was the placement of some rarely sampled taxa. Taxa that are present only in some publications have been allocated a high proportion of missing character values, which makes MP analyses much more difficult. It is difficult to know how many of these taxa are misplaced because no reference phylogeny for the 401 genera is available, but tests of MRP reconstruction done without constraining three clades (panicoids, pooids and chloridoids) to be monophyletic ended with rarely sampled taxa scattered all around the supertrees (data not shown). This can be an important problem when MRP analyses are performed on a wide taxonomic group comprising large numbers of rarely sampled taxa.

MRP coding, especially when weighted by node support, can be considered as an indication of the signal in the primary data, with each node represented by one (weighted) synapomorphy (see Bininda-Emonds et al., 1999). However, supertrees cannot be viewed and interpreted in exactly the same way as phylogenetic analyses based on biological characters. One major problem is the difficulty of assigning node support for supertrees. Bootstrapping and other resampling procedures cannot be applied due to the clear non-independence of the characters present in the binary matrix. It should however be possible to resample the source trees rather than each individual character. This strategy would therefore bootstrap or jackknife blocks of binary characters that define the source trees. Bremer support (Bremer, 1988) is also often used with MRP, but this method is certainly equally affected by this problem of non-independence of characters (as MP itself is, hence the different MRP modifications proposed). Moreover, branch lengths associated with supertrees are treated in the same way as morphological branch length, and unlike molecular branch length, it is not possible to use them for establishing divergence times. This weakness is even more significant in taxonomic groups, such as the grasses, where dates are difficult to gather from other sources such as fossil records (but see chapter 6).

4.5 Conclusion

Supertrees offer an easy way of producing phylogenetic trees with a high number of taxa and these can give good estimates of relationships within these groups (in this case the grass family). Supertrees using the eight character partitions of the GPWG data sets were found to roughly match the combined analysis, the discrepancies being found for the best MRP modifications mainly in weakly supported branches of the combined tree. The Baum and Ragan and the Purvis modifications were found to give similar results, while incorporating bootstrap support associated with pre-existing topologies improved the Baum and Ragan modification. Moreover, supertrees can be useful for comparative studies (Purvis et al., 1995; Bininda-Emonds et al., 1999), whether of adaptation, co-evolution, rates of evolution, co-speciation or rates of effective cladogenesis, where accuracy in the branch length of phylogenetic reconstructions is not the primary problem (Purvis et al., 1994). Supertree reconstructions are also a useful way to help highlight poor taxonomic sampling and identify where previous studies are inconsistent. They can therefore be used as an exploratory tool capable of developing new hypotheses and indicating where future research should be focused.

The various large supertrees built for the grass family have also highlighted another significant issue in grass phylogenetics. Indeed, it is unclear how some of the major clades or subfamilies of grasses are interrelated, and different evolutionary hypotheses have been proposed depending on the data sets used or the method selected to perform the analysis. In particular, the relationships between the clades containing wheat, rice, and maize are still unclear and an assessment of the current hypotheses is presented in the next chapter.

CHAPTER 5. IS WHEAT SISTER TO RICE OR TO MAIZE : ERROR, BIAS AND INCONSISTENCY IN GRASS PHYLOGENETICS

5.1. Introduction

Accurate grass phylogenetic trees can be used as important tools to better understand character and genome evolution, and to test a wide range of evolutionary hypotheses (Felsenstein, 1985b; Harvey and Pagel, 1991; Pagel, 1999). Recent evidence of micro- and macro-synteny among genes within the family (Bennetzen and Freeling, 1993; Bennetzen, 2000; Chandler and Wessler, 2001) also indicate that grasses may be viewed as a general genetic system where gene order and quantitative trait loci (QTL) could be predicted for all grass species (Freeling, 2001; Jones et al., 2002) by using the relevant information from only a few well characterized species (e.g. rice, maize). However, small-scale genome rearrangements and deletions have complicated the microlinearity between closely related grass species (e.g., sugarcane and maize), but also between rice and other crop plants (Keller and Feuillet, 2000). One potential cause of these rearrangements could be artificial selection, but understanding these rearrangements and deletions in an evolutionary framework will allow a better characterization of grass genomes and improve their utilization.

5.1.1 Grass phylogenetics

Recent molecular analyses have modified our understanding of grass evolutionary history, and in particular the relationships between the PACCAD clade and the Bambusoideae, Ehrhartoideae and Pooideae. A majority of gene trees representing DNA regions from the plastid and the nuclear genomes support the BE-P clade. Analyses based on *gbss* (Mason-Gamer et al., 1998), *phyB* (Mathews and Sharrock, 1996; Mathews et al, 2000), *rpl16* (Zhang, 2000), *ndhF* (Clarke et al., 1995) all resolved the BE-P clade with varying degrees of support. However, an alternative hypothesis to the BE-P clade is supported by another set of analyses, which are based on DNA regions from the nuclear (*Adh* – Gaut et al., 1999; ITS – Hsiao et al., 1999) as well as the plastid genome (*rbcL* – Duvall and Morton, 1996; Gaut et al., 1997; *rpoC2* – Cummings et al., 1994; Barker et al., 1999). Instead of placing pooids within bambusoids and ehrhartoids (i.e. the BE-P clade), these gene trees show pooids as sister to the PACCAD clade (i.e. the PACCAD-P clade), and thus within a larger clade that excluded bambusoids and ehrhartoids. However, Barker et al. (1999) noted that topologies obtained with *rpoC2* could support both hypotheses (BE-P or PACCAD-P) depending on the coding of indels during the analysis. Although the combined analysis of six different DNA regions, plastid DNA restriction sites, and morphological characters resulted in the recognition of the BE-P clade, bootstrap percentages for the group remain relatively low (71%; GPWG, 2001). Addition of other molecular data sets that do not support the BE-P clade (such as *Adh*) could change the result obtained. Finally, another approach using supertree reconstruction based on combination of 55 published source trees for grasses resulted in pooids being sister to the PACCAD clade (Salamin et al., 2002).

5.1.2 Error, bias and inconsistency

A major concern that arises is whether the gene trees truly reflect the organismal phylogeny (Brower et al., 1996; Doyle, 1997; Maddison, 1997; Page and Charleston, 1997; Giannasi et al., 2001). Gene trees and organismal phylogenies can differ because of retention of ancestral polymorphisms, reticulation among populations or species, or rapid diversification (Wendel and Doyle, 1998). This is of particular concern for the non-recombining organelle genome because the effects of reticulation are potentially retained through subsequent generations (e.g. Hodkinson et al., 2002a). Beside these biological reasons to explain incongruences between gene trees, it is also important to question other artefacts that could affect the accuracy of phylogenetic reconstructions, such as errors in phylogenetic reconstructions (Felsenstein, 1978a; Hillis and Huelsenbeck, 1994; Swofford et al., 1996). Taxa sampling in all molecular studies involving grasses have been limited to a small portion of the diversity in the family and the number of characters considered remains low, typically less than 2,000 bp. Even if evolution occurred exactly as assumed by a particular analytical method, an incorrect tree may be inferred with finite data due to chance events alone, which introduces *random error*. When evolutionary processes violate the assumptions of a phylogenetic method, systematic error will arise (Felsenstein, 1978a; Hillis and Huelsenbeck, 1994; Swofford et al., 1996). Under these conditions, mistaken inferences can be more or less random, or, if only certain incorrect topologies are preferred, they can be *biased* in the context of the underlying process of molecular evolution for those taxa. Because the effect is systematic, the addition of more data will tend to solidify the incorrect solution and the method is said to be *inconsistent* (Felsenstein, 1978a). This situation is well known with MP that has been shown to be inconsistent when dealing with simple trees (Felsenstein, 1978a; Hendy and Penny, 1989), but ML and distance methods can also be inconsistent when the assumed model of evolution is incorrect (Kuhner and Felsenstein, 1994; Chang, 1996).

Several methods have been proposed to identify random error, bias, and inconsistency in real data (Huelsenbeck, 1997; Lyons-Weiler and Hoelzer, 1997). An approach is to use Monte Carlo simulations to examine whether a phylogenetic reconstruction method is biased under some model conditions estimated from a given data set. Several groups of plant and animals, showing putative problematic long-branch attraction, have been examined using Monte Carlo simulations, and results have shown that branches could in some case be long enough to cause wrong topologies to be reconstructed with high probabilities (Huelsenbeck et al., 1996; Huelsenbeck, 1998; Maddison et al., 1999; Pellmyr and Leebens-Mack, 1999). In particular, Sanderson et al. (2000) used an extensive series of simulations to investigate the effect of error and bias in land plant phylogenetics, and the position of the Gnetales in particular.

5.1.3 Aims

In this chapter, a similar approach to Sanderson et al. (2000) was used to investigate the possible occurrence of error, bias and inconsistency in grass phylogenetic reconstructions based on six different DNA regions sequenced for 66 species. Our goal was to assess whether random or systematic error could explain the disparate but strongly supported phylogenetic hypotheses that are obtained for the grass family with different genes, or if it is necessary to seek other explanations for conflict between these molecular data. At the same time, although third codon positions have been demonstrated to contain more phylogenetic signal (see chapter 2), the first and second codon positions were used to investigate the effect of evolutionary rates on the amount of error and bias in grass phylogenetics. First, a combined DNA matrix from the GPWG was re-analysed using MP, ML, and NJ to evaluate if more consistent methods such as ML would resolve the incongruence between gene trees. Bayesian inference was also used to obtain posterior probabilities for the competing evolutionary hypotheses for the grasses. Finally, Monte Carlo simulations were performed to assess the potential effects and extent of random error and bias on grass phylogenetic reconstructions.

5.2. Material and Methods

5.2.1 Data partitions

The data matrix used was taken from the GPWG (GPWG, 2001). The matrix of six DNA regions includes 66 grass species representing all the currently recognised subfamilies. Not all species had been sequenced for each DNA region, and there were, therefore, several missing data. Three plastid regions, *ndhF*, *rbcL*, and *rpoC2*, and three nuclear regions, *phyB*, *gbss*, and the nuclear ribosomal region (rDNA) comprising the 5.8S rDNA and the internal transcribed spacer 2 (ITS2) were re-analysed separately, altogether or in plastid versus nuclear partitions. Exons (coding sequences) were also analysed using the three codon positions together (hereafter referred to as 1+2+3) and two partitions of the data, the first and second codon positions (hereafter referred to as 1+2), or the third codon positions alone (hereafter referred to as 3). Furthermore, rDNA sequences were analysed using 5.8S and ITS2 together or using 5.8S and ITS2 regions separately. Partition homogeneity tests (PHT; Farris et al., 1995) as implemented in PAUP*4b were conducted to test whether significant conflicting signals were present between partitions for each DNA regions and between those regions.

5.2.2 Phylogenetic analyses

Trees were reconstructed using MP, NJ and ML as implemented in PAUP*4b. Taxa analysed were those that have been sequenced for the particular DNA region or set of DNA regions forming the combined data sets. The only exception was for all DNA regions combined, where all taxa having more than four DNA sequences were included, otherwise only four taxa could have been used in the analyses. MP analyses used heuristic search options consisting of 500 random addition sequences keeping up to 20 trees followed by TBR branch-swapping until completion.

NJ analyses were performed using Log-Det or paralinear distance transformation (Lockhart et al., 1994) and HKY85+Γ distances (Hasegawa et al., 1985). The different possible models of DNA evolution used with ML were tested using ModelTest3.6 (Posada and Crandall, 1998). The GTR+I+Γ substitution model (Rodriguez et al., 1990) best fitted *ndhF*, *rpoC2*, *phyB*, *gbss*, ITS2+5.8S, and combined plastid and nuclear sequences, while the TrN+I+Γ substitution model

(Tamura and Nei, 1993) was selected for *rbcL* and all six partitions combined. These two models were therefore used in subsequent ML analyses. Because of the computational burden associated with ML, the model parameters were estimated on a fixed topology given by the NJ trees for each data partition. These parameters were then fixed in the ML searches and empirical base frequencies were used. Heuristic searches were performed using NJ to obtain a starting tree and then followed by NNI without saving multiple trees. Bootstrap percentages were calculated using 500 replicates and the same heuristic search options as described above for MP, and 500 replicates for NJ with both Log-Det and HKY85+Γ distances. Bootstrap percentages were not calculated for ML due to the time burden implied.

SH test (Shimodaira and Hasegawa, 1999; Goldman, et al., 2000) as implemented in PAUP*4.0b were performed to evaluate if one of two alternative hypotheses for grass evolution (Fig. 5.1) was explaining the different data partitions significantly better than the other. The tests were carried out using the RELL estimation (Shimodaira and Hasegawa, 1999) to speed up the procedure. In effect, the RELL estimation fixes the parameters of the model of substitution at the values obtained from the observed character matrix, and uses these estimates for all bootstrap replicates instead of re-estimating the parameters for each bootstrap replicate.

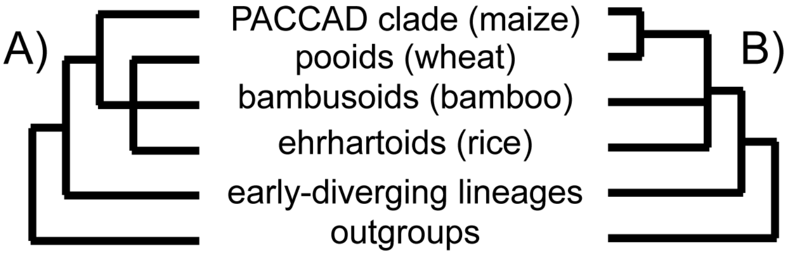


Figure 5.1 - Two different evolutionary hypotheses concerning the deep branching pattern in the grass family. A) The BE-P hypothesis where pooids are sister to bambusoids and ehrhartoids. B) The PACCAD-P hypothesis where pooids are sister to the PACCAD clade.

Bayesian analyses were also conducted on each data partition using MrBayes (Huelsenbeck and Ronquist, 2001). Due to computer memory limitation (Pentium III 600MHz microprocessor, 256Mb RAM, running LINUX SuSE7.3), between 18 and 20 species were sampled from the original matrix for each data partition. The sampled species were selected in order to reduce the amount of missing data, but when a species could not be selected for a DNA region, another representative of the same subfamily was chosen to maintain sample size within all major subfamilies. The HKY85+Γ model of substitution was chosen for the Bayesian analyses, with parameters fixed to the same values as ML analyses. 500,000 generations were performed on four Markov chains with trees sampled every hundred generations, with a uniform prior distribution. The trees obtained from the first 25,000 generations were discarded to allow for the burn-in of the process, as the likelihood function stabilised shortly before reaching 25,000 generations in all MCMC searches.

5.2.3 Evaluating error, bias and inconsistency

Computer simulations were used to evaluate random error, bias, and statistical inconsistency in the different partitions of the GPWG matrix. Two different model trees, corresponding to the two major evolutionary hypotheses for the deep branching pattern within the grass family (Fig. 5.1), were constructed based on the literature and our re-analyses of the GPWG matrix. The BE-P model (Fig. 5.1A) places the pooids in a monophyletic clade containing the bambusoids and the ehrhartoids, while the PACCAD-P model (Fig. 5.1B) creates a clade containing the pooids and the six subfamilies composing the PACCAD clade. Thus, the BE-P model groups rice with wheat and the PACCAD-P model maize with wheat. The two contradictory topologies were successively enforced in MP searches using the same heuristic options as described above.

Branch lengths and parameters of the HKY85+Γ substitution model were then estimated under ML based on each character partition, except for the base frequencies, which were fixed at their empirical values. Only one set of parameters was used for the two model trees simulated (Table 5.1). These topologies with branch lengths and ML estimates of the model parameters were used to simulate 500 new data sets for the character partitions examined using the program evolver from the PAML3.1 package. MP searches were then performed on each new data set using heuristic search options consisting of 100 replicates of random addition sequence keeping up to 50 trees with TBR branch-swapping. The saved trees were then compared to the model trees to calculate the percentages of correct trees recovered by MP using the program TreeCorrect1.2b[8]. A saved tree was considered ‘correct’ if its topology was compatible with the model of evolutionary hypothesis under consideration (i.e. BE-P or PACCAD-P; Fig. 5.1). To evaluate inconsistency in phylogenetic reconstructions, the simulations were repeated by keeping the conditions of the simulations identical to those described above (i.e. branch lengths, model parameters, topology), but we increased the size of each data set to 5,000 and 10,000 bp instead (Sanderson et al., 2000).

The number of taxa was also increased for some data partitions that showed potential long-branch attraction problems by repeatedly breaking the longest branches in half and adding new taxa until model trees of 150 and 500 taxa were created, using the software BranchCut[9].

An outline of the different analyses performed on the GPWG data set is presented in Fig. 5.2.

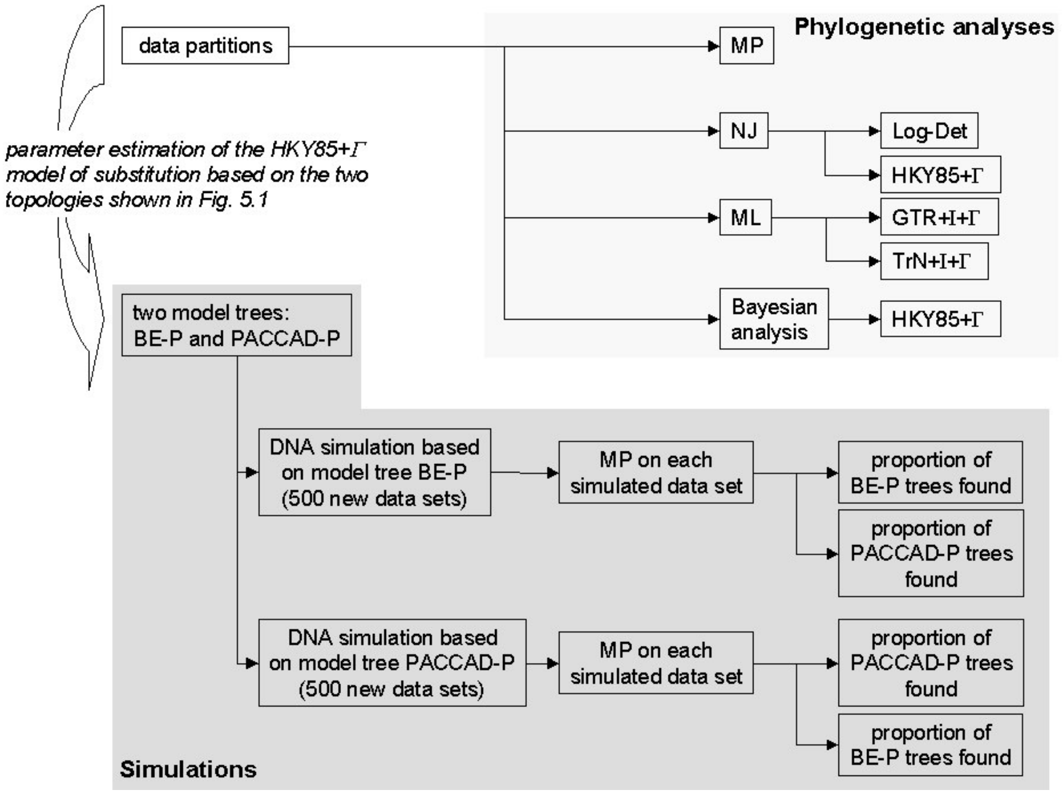

data partitions
Phylogenetic analyses
MP
NJ
Log-Det
HKY85+Γ
ML
GTR+I+Γ
TrN+I+Γ
Bayesian analysis
HKY85+Γ
parameter estimation of the HKY85+Γ model of substitution based on the two topologies shown in Fig. 5.1
two model trees: BE-P and PACCAD-P
DNA simulation based on model tree BE-P (500 new data sets)
MP on each simulated data set
proportion of BE-P trees found
proportion of PACCAD-P trees found
DNA simulation based on model tree PACCAD-P (500 new data sets)
MP on each simulated data set
proportion of PACCAD-P trees found
proportion of BE-P trees found
Simulations

Figure 5.2 - Schematic representation of the analyses performed to evaluate the error and bias in grass phylogenetics.

[8] see Chapter 8: Technical notes for a description

[9] See Chapter 8. Technical notes for a description

Table 5.1 - ML estimates of the HKY85+Γ substitution model parameters calculated on the respective MP trees and used to simulate the DNA sequences.

Sequences	base pairs	kappa	alpha	Nucleotide frequencies			
				A	C	G	T
all	6610	3.634	0.261	0.265	0.203	0.238	0.293
plastid	4332	4.466	0.308	0.287	0.171	0.209	0.332
nuclear	2278	3.609	0.322	0.209	0.284	0.309	0.196
ndhF 1+2+3	2211	4.196	0.381	0.273	0.164	0.175	0.386
ndhF 1+2	1474	3.121	0.279	0.262	0.188	0.196	0.353
ndhF 3	737	5.991	0.888	0.294	0.117	0.133	0.454
rbcL 1+2+3	1344	3.557	0.202	0.271	0.192	0.249	0.287
rbcL 1+2	896	1.286	0.014	0.263	0.211	0.298	0.227
rbcL 3	448	6.366	0.787	0.286	0.156	0.151	0.406
rpoC2 1+2+3	777	5.069	0.981	0.405	0.146	0.281	0.166
rpoC2 1+2	518	4.453	0.881	0.405	0.122	0.324	0.147
rpoC2 3	259	6.186	2.982	0.405	0.193	0.195	0.205
phyB 1+2+3	1182	4.273	0.353	0.215	0.269	0.292	0.223
phyB 1+2	788	2.281	0.321	0.249	0.255	0.269	0.225
phyB 3	394	5.786	1.724	0.147	0.297	0.336	0.218
gbss 1+2+3	774	2.714	0.615	0.213	0.294	0.338	0.155
gbss 1+2	516	3.224	0.252	0.295	0.206	0.301	0.198
gbss 3	258	5.282	1.354	0.049	0.471	0.412	0.068
ITS	322	3.781	0.301	0.187	0.316	0.331	0.165
5.8S	163	4.886	0.108	0.228	0.294	0.285	0.192
ITS2	159	4.126	0.781	0.142	0.341	0.381	0.136

5.3 Results

5.3.1 Phylogenetic reconstructions

PHT indicated no significantly different signals at the 5% level between 1+2 and 3 codon partitions for *ndhF*, *rpoC2*, *phyB*, and between 5.8S and ITS2 (*p*=0.483, *p*=0.164, *p*=0.481 and *p*=0.981, respectively), while *rbcL* and *gbss* had significantly different signal between 1+2 and 3 codon partitions (*p*=0.025 and *p*=0.012, respectively). Comparisons between DNA regions suggested that no conflict was present between *ndhF* / *rpoC2*, *ndhF* / *phyB*, *rpoC2* / *phyB*, *rpoC2* / *gbss*, and *gbss* / ITS2+5.8S (*p*=0.152, *p*=0.174, *p*=0.806, *p*=0.981 and *p*=0.973, respectively). All other comparisons resulted in significant conflict (*p*<0.05).

There was little contradiction between the evolutionary hypotheses resulting from MP, ML, or NJ analyses (Table 5.2). Phylogenetic searches based on *ndhF, phyB, gbss*, nuclear and all partitions combined, resulted in the placement of the pooids within the BE-P clade, while those based on *rbcL* always placed the pooids as sister to the PACCAD clade (i.e. PACCAD-P hypothesis; Table 5.2 and Fig. 5.1). It is noteworthy that while the BE-P clade was supported by phylogenetic analyses of *gbss*, PACCAD was not monophyletic in all analyses based on this nuclear region. MP and ML recovered the PACCAD-P clade with ITS2+5.8S and plastid sequences, while NJ based on Log-Det or HKY85+Γ distances supported the BE-P clade. In ML analyses, *rpoC2* resolved the BE-P clade, while MP and NJ placed the PACCAD clade sister to the bambusoids and ehrhartoids. For all data partitions and all methods used, the nodes supporting the BE-P or PACCAD-P clades received low bootstrap percentages usually under 80%, except with *rbcL*, *phyB* for MP and NJ and for the plastid and all partitions combined for NJ where percentages were higher or equal to 80% (Table 5.2). SH tests (Table 5.3) showed no significant difference between the two alternative hypotheses, except for *phyB* (*p*=0.021).

Table 5.2 - Evolutionary hypotheses obtained with MP, ML (model = GTR+I+Γ or TrN+I+Γ), and NJ (distance = Log-Det and HKY85+Γ) algorithms. Numbers in parentheses are bootstrap percentages supporting each hypothesis. With NJ, the first figure indicates values obtained with Log-Det distance, and the second one with HKY85+Γ distance.

Algorithm	Evolutionary hypotheses		
	BE-P		PACCAD-P
MP	*ndhF* (52)	*rpoC2* (67)	*rbcL* (79)
	gbss[1] (39)	*phyB* (90)	ITS (14)
	nuclear (56)	all (64)	plastid (83)
ML	*ndhF*	*rpoC2*	*rbcL*
	gbss	*phyB*	ITS
	nuclear	all	plastid
NJ	*ndhF* (78/72)	*rpoC2*[2] (34/39)	*rbcL* (92/93)
	gbss (61/55)	*phyB* (88/85)	
	ITS (16/22)	nuclear (50/35)	
	plastid (71/77)	all (98/100)	

[1] BE-P clade, but PACCAD polyphyletic

[2] PACCAD clade sister to bambusoids+ehrhartoids with HKY85+Γ distance only

Table 5.3 - SH tests using RELL estimation procedure between two alternative hypotheses for the grasses.

Genome	Sequences	P values
	all	0.383
	plastid	0.485
	nuclear	0.404
plastid	*ndhF*	0.433
	rbcL	0.419
	rpoC2	0.192
nuclear	*phyB*	0.021
	gbss	0.213
	ITS	0.171

The use of Bayesian methods allowed us to obtain posterior probabilities for each clade conditioned over the different data set partitions (Table 5.4). A high probability of 0.947 for the PACCAD-P clade was associated with the combined data set, while a maximum posterior probability of one for the BE-P clade was associated with the combined nuclear sequences (Table 5.4). The combined plastid sequences favoured the PACCAD-P clade, although the probability was low (0.389; Table 5.4). For the plastid sequences, the BE-P clade obtained the highest posterior probability only with *ndhF* 1+2+3 and *ndhF* 1+2 (0.275 and 0.449 respectively; Table 5.4). The PACCAD-P clade obtained either the highest posterior probabilities with *ndhF* 3 and *rbcL* 3 or was the only clade considered in the study that was found during the searches with *rbcL* 1+2+3 and *rbcL* 1+2 and with all partitions of *rpoC2* (Table 5.4). However, the probabilities obtained for *rbcL* 1+2 and *rpoC2* 3 were low (0.085 and 0.071 respectively; Table 5.4). For the nuclear regions, the BE-P clade obtained the highest posterior probabilities with all partitions of *phyB* (1+2+3, 0.984; 1+2, 0.389; 3, 0.364; Table 5.4) and with *gbss* 1+2 (0.907; Table 5.4), while the PACCAD-P clade was marginally preferred with ITS2+5.8S and ITS2 alone (0.539 and 0.416 respectively; Table 5.4). Finally, *gbss* 1+2+3 and *gbss* 3 did not support either the BE-P or the PACCAD-P clade and 5.8S gave probabilities lower or equal to 0.005 for both clades (Table 5.4).

Table 5.4 - Posterior probabilities for two evolutionary hypotheses.

Genome	Sequences	Evolutionary hypotheses	
		BE-P	PACCAD-P
	all	0.015	0.947
	plastid	0.092	0.389
	nuclear	1.000	0.003
plastid	*ndhF* 1+2+3	0.275	0.050
	ndhF 1+2	0.449	0.010
	ndhF 3	0.083	0.145
	rbcL 1+2+3	-	0.969
	rbcL 1+2	-	0.085
	rbcL 3	0.015	0.562
	rpoC2 1+2+3	-	0.997
	rpoC2 1+2	-	1.000

	rpoC2 3	-	0.071
nuclear	*phyB* 1+2+3	0.984	-
	phyB 1+2	0.389	0.008
	phyB 3	0.364	0.011
	gbss 1+2+3	-	-
	gbss 1+2	0.907	0.050
	gbss 3	-	-
	ITS	0.172	0.539
	5.8S	0.005	0.003
	ITS2	0.387	0.416

Table 5.4, continued

5.3.2 Simulations

Figure 5.3 shows the ML estimates of branch length for the six DNA regions investigated. Branch lengths in those trees vary greatly, but most DNA sequences produced trees characterized by long branches leading to the terminal taxa in comparison to internal branches separating the major subfamilies (Fig. 5.3). The stem lineage subtending the PACCAD clade and the pooids was relatively long in comparison to the internal branches leading to the other subfamilies with all DNA regions (from 5 to 20 times longer; Fig. 5.3), except *gbss* where PACCAD was not monophyletic. Branches subtending the BE-P or PACCAD-P clades were short (between 0.0009 for BE-P with *ndhF* and 0.0111 for PACCAD with *phyB*; Fig. 5.3) , especially with *ndhF*, *rpoC2,* and ITS2+5.8S. These branches were longer and the groups they subtended were better defined with *rbcL* and *phyB*.

The percentage of MP trees correctly recovering the BE-P or PACCAD-P clades was recorded for all simulated data sets based on both model trees (Fig. 5.4). Results differed substantially between DNA regions and genomes. Combining the plastid, the nuclear, or all six partitions reduced greatly the error in phylogenetic reconstruction. The model tree was recovered in more than 75.5% of the combined reconstructions, except for the PACCAD-P model tree with simulated data sets based on the combined nuclear DNA where only 50.8% of the MP trees were correct (Fig. 5.4). However, in each case no bias toward the alternative evolutionary hypothesis could be found, except for the plastid sequences when BE-P was the true tree where low bias was found towards the PACCAD-P model tree.

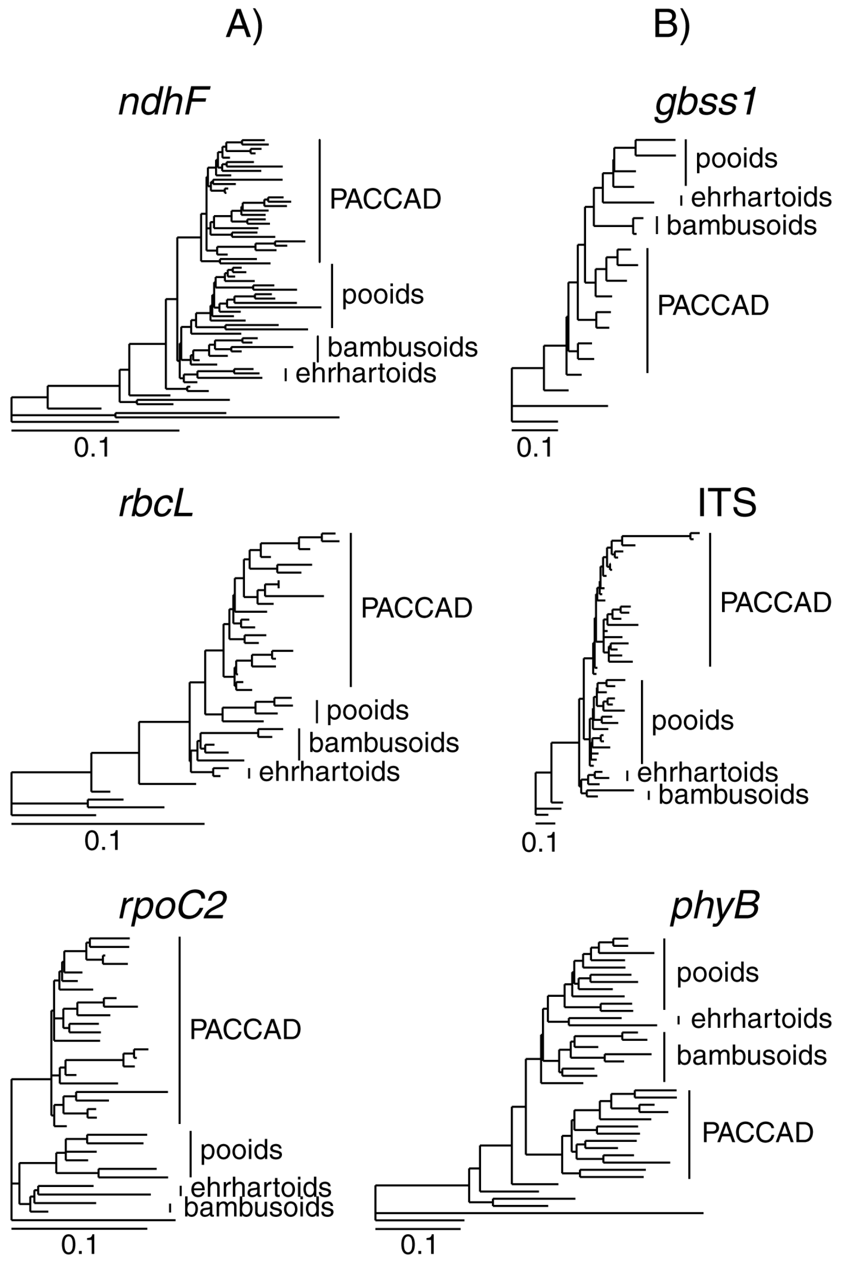


*Figure 5.3 - ML branch length estimates for A) three plastid DNA sequences (*ndhF*,* rbcL*,* rpoC2*), B) three nuclear DNA sequences (*phyB*,* gbss*, ITS).*

When looking at each DNA region separately, MP was biased with a 30 to 35% chance of obtaining the BE-P clade even when the model tree included the PACCAD-P clade for *ndhF* 1+2 and 1+2+3. For *phyB* 3 and 1+2+3, there was 17 and 18% chance respectively of mistakenly obtaining the BE-P clade when the PACCAD-P clade was the model. The bias was smaller for *ndhF* 3 and *phyB* 1+2 although MP could not retrieve with high probability the true PACCAD-P model tree in any partitions of either *ndhF* or *phyB*. For *rbcL*, MP was not able to recover the correct tree more than 60% of the time with any model tree and any codon partition and *rbcL* 1+2 had ca. 20% chance of finding the PACCAD-P clade when the model tree included the BE-P clade. For *rpoC2* 1+2+3, MP had more than a 75% chance of recovering both correct trees, while MP had difficulties in building the PACCAD-P model tree with *rpoC2* 3 and the BE-P model tree with *rpoC2* 1+2. The latter codon partition introduced a strong bias in the analysis, forcing MP towards the PACCAD-P clade (29% chance) when the BE-P clade was the model (Fig. 5.4). For ITS2+5.8S, the correct BE-P or PACCAD-P tree was found 49% and 72% of the times respectively, with a bias towards the PACCAD-P clade when the BE-P model tree was used to simulate the data. More or less the same pattern as ITS2+5.8S was followed by ITS2 with a slightly stronger bias and lower probabilities of obtaining both model trees, while MP only recovered the BE-P tree 25% of the time and never retrieved the PACCAD-P model tree. The *gbss* results also showed low probabilities to recover the BE-P tree and slightly higher probabilities to obtain the PACCAD-P tree, except for *gbss* 1+2 where MP had 80% chance to recover the PACCAD-P tree.

The results from increasing the number of characters in each simulated data set to 5,000 and 10,000 bp are shown in Figure 5.5; this time only all codon partitions combined were considered. For most DNA regions the error rate was reduced and MP had more than 72% chance of finding the correct trees (i.e. BE-P or PACCAD-P) when *rbcL*, *rpoC2*, ITS2+5.8S, plastid, nuclear and all DNA regions combined were used. MP searches based on simulated *rbcL* data sets needed more than 10,000 bp to find the true trees with a higher probability than 93%, while MP found the true tree 80% of the time when simulations were based either on combined nuclear data and the PACCAD-P model tree or on all six sequences combined and the BEP model tree.

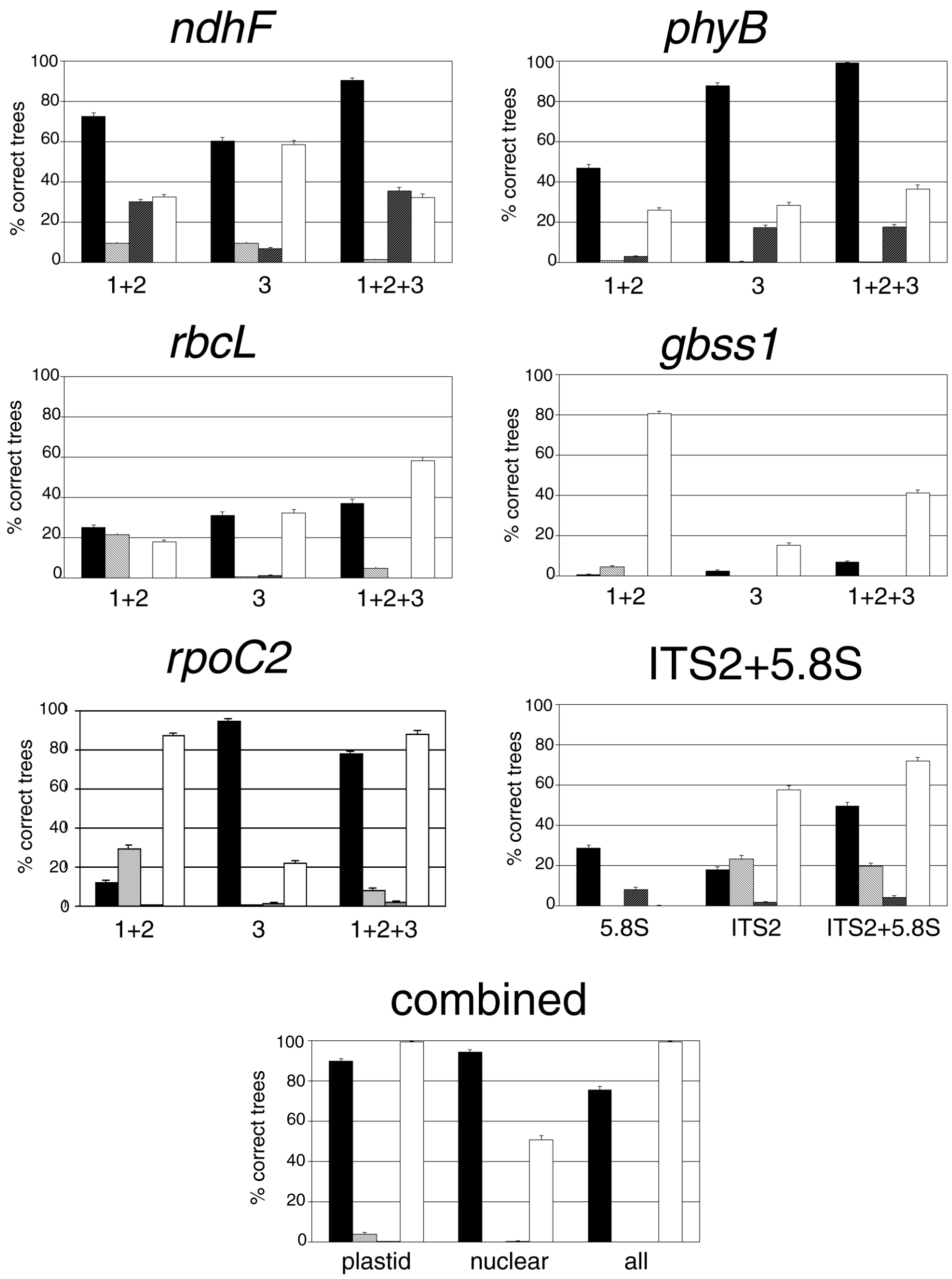


Figure 5.4 - Estimated error rates and standard error for MP calculated from Monte Carlo simulations given the six DNA regions examined separately or in combination. Black bars indicate percentages of trees retrieving BE-P when the BE-P model was indeed true, white bars indicate the percentages of trees retrieving PACCAD-P when the PACCAD-P model was indeed true, light grey bars indicate percentages of trees retrieving PACCAD-P when BE-P model was indeed true and dark grey bars indicate percentage of trees retrieving BE-P when PACCAD-P was indeed true.

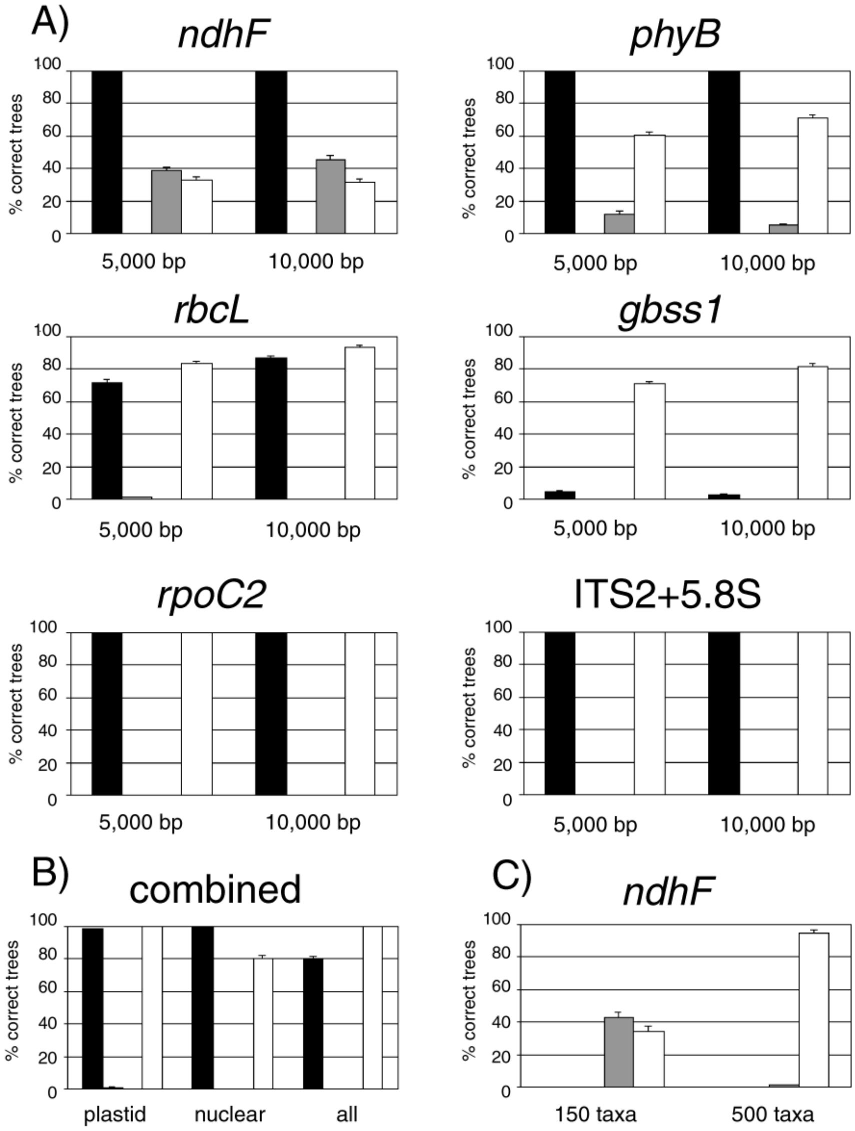


Figure 5.5 - Estimated error rates and standard error for MP calculated from Monte Carlo simulations given A) six DNA regions examined separately with increasing number of characters, B) combined DNA regions for 10,000 bp only, and C) ndhF *with increasing number of taxa. Black bars indicate percentages of trees retrieving BE-P when the BE-P model was indeed true, white bars indicate the percentages of trees retrieving PACCAD-P when the PACCAD-P model was indeed true, light grey bars indicate percentages of trees retrieving PACCAD-P when BE-P model was indeed true and dark grey bars indicate percentage of trees retrieving BE-P when PACCAD-P was indeed tree.*

The bias toward the BE-P model tree, when PACCAD-P tree is correct, increased for *ndhF* with the increase in the number of characters (35% with 2,210 bp; 39% with 5,000 bp; 46% with 10,000 bp; Figs. 5.4 and 5.5), with an associated decrease in the percentages of correct trees found. Increasing the number of taxa to 150 instead of 65 by breaking the longest branches in the PACCAD-P model tree did not alter the bias found for *ndhF* (Fig. 5.5C). However, by continuing to break the longest branches until 500 taxa were reached considerably reduced the bias towards the BE-P model tree (42% and 1% for 150 and 500 taxa respectively; Fig. 5.5C). In contrast, for *phyB*, the same bias decreased as the number of characters available for the analysis increased (18% with 1182; 12% with 5,000 bp; 5% with 10,000 bp; Figs. 5.4 and 5.5). Finally, data sets simulated based on *gbss* and the BE-P model tree proved more and more difficult for MP when the number of characters was increased, and this was in contrast to the PACCAD-P model tree which was more and more often recovered with character sets of 5,000 and 10,000 bp (71% and 82% respectively; Fig. 5.5).

5.4 Discussion

5.4.1 Phylogenetic methods

The methods of phylogenetic reconstruction used on the different partitions gave a consistently different picture for grass subfamilial relationships, whether MP, ML, or NJ searches were performed. The conflict between the different DNA regions observed with MP is therefore still present even when potentially more robust methods are used and it is not possible to assert whether random or systematic errors are the reason for these discrepancies. The comparison was restricted to the placement of the major subfamilies and to the BE-P vs. PACCAD-P hypotheses in particular (Fig. 5.1). Two out of three plastid regions (*ndhF* and *rpoC2*), representing more than two thirds of the total number of characters and more than three quarters of the number of parsimony-informative characters, supported the BE-P clade. However, the combined plastid sequences favoured PACCAD-P (Table 5.1). Bootstrap support was higher using *rbcL* than *ndhF* or *rpoC2* (79% vs. 52% and 67% respectively; Table 5.1), and it could explain the result found with the combined plastid sequences. Our results for plastid sequences differed from the GPWG analyses (GPWG, 2001) because only taxa that had sequences for the three plastid sequences in the GPWG matrix were included. The *ndhF* region has been sequenced for 65 out of 66 taxa, while only 37 and 34 taxa were sequenced for *rbcL* and *rpoC2*. The removal of these taxa with large amounts of missing values seems therefore to have an impact on the result of the plastid analysis (see Wilkinson, 1995; Wiens, 1998; Anderson, 2001 for a discussion on missing values).

Using a method that is potentially statistically consistent such as ML, could provide a way to tease apart random error, bias and inconsistency in grass phylogenetic studies (Huelsenbeck, 1997). ML estimates are not guaranteed to be unbiased (Lehmann, 1983), and under some model conditions, likelihood methods are not as efficient as nonparametric methods such as MP (Huelsenbeck, 1998; Siddall, 1998). Swofford et al. (2001) have also shown that reasons for the efficiency of MP in long-branch repulsion problems are due to bias. Another potential problem affecting our ML searches is the type of heuristic search used. However, due to the computational and time burden implied with ML, it was not possible to spend more time in performing these searches (depending on the data partition analysed, between 14 to 24 h of CPU time was spent for one heuristic search with a NJ starting tree, no multiple trees saved and NNI swapping algorithm on a Pentium III 600MHz microprocessor, 256 Mb RAM, running LINUX SuSE7.3). NJ analyses were performed with the Log-Det distance, which may be relatively robust to composition changes across the tree (Lockhart et al., 1994). Trees obtained from ITS2+5.8S and plastid sequences shifted from PACCAD-P to BE-P model tree with the Log-Det distance, thus potentially indicating shifts in base composition across taxa. The Log-Det distance does not allow for site-to-site variation in rates because of theoretical problems arising with non-stationary base composition and site-to-site rate variation (Bakke and von Haeseler, 1999). The rate heterogeneity among sites is pervasive in the different data partitions (Table 5.1) and this may have an impact on the phylogenetic analyses. However, introducing a Gamma distribution with the HKY85 distance did not change the results either (Table 5.1). The differences obtained between the different methods of reconstruction are more likely to be due to other factors (e.g. taxon sampling, size of the DNA sequences) rather than to changes in nucleotide frequencies across the grass species or to rate heterogeneity among sites.

5.4.2 Bayesian inference

A Bayesian approach to phylogeny reconstruction expresses the uncertainty in the phylogeny and in the parameters of the sequence substitution model with a posterior probability distribution. One advantage of Bayesian phylogenetic inference is that the posterior probabilities of all trees will sum up to one, and hence competing evolutionary hypotheses can be compared through their posterior probabilities (Yang and Rannala, 1997; Larget and Simon, 1999). However, it was difficult to draw a clear pattern from the Bayesian analyses, the results being sometimes very different from the other phylogenetic methods. Contrary to the study of Whittingham et al. (2002), bootstrap values obtained with MP and posterior probabilities obtained from the Bayesian analysis do not give similar results and it was difficult to find a common pattern. Some data partitions producing high posterior probabilities for one hypothesis were associated with low bootstrap percentages for the alternative hypothesis (e.g. nuclear, all or *rpoC2*; Tables 5.1 and 5.3) or low posterior probabilities with high bootstrap values (e.g. plastid; Tables 5.1 and 5.3). At the same time, some analyses received high bootstrap values and high posterior probabilities for the same evolutionary hypothesis (e.g. *rbcL*, *phyB*; Tables 5.1 and 5.3) or low bootstrap values and low posterior probabilities (e.g. *ndhF*, *gbss*, ITS2+5.8S; Tables 5.1 and 5.3) for the same evolutionary hypothesis. A similar conclusion arises by comparing the results of the SH tests and Bayesian analyses. Although the two alternative hypotheses were most of the time indistinguishable using the SH tests, Bayesian analysis was most of the time in clear favour of one or the other of the hypotheses (Tables 5.2 and 5.3).

5.4.3 Error and bias

The former approach to resolve error in phylogenetic inference is impaired by the fact that the true tree cannot be known. Although using a consistent and robust method of phylogenetic inference should reduce the impact of random or systematic error, it is impossible to totally exclude such error. In real data sets, even if ML is consistent and MP not, the issue is the extent of the bias, rather than the asymptotic behaviour as more data are obtained. In contrast, Monte Carlo simulations allow the definition of model tree and a clear specification of the parameters introduced in the simulations. It is therefore a powerful approach to determine whether phylogenetic methods are affected by errors, and to what extent (Huelsenbeck, 1998; Sanderson et al., 2000). One of the key results highlighted in our results is the lack of phylogenetic information contained in some of the plastid or nuclear sequences analysed separately (Fig. 5.4).

It is also apparent that in most cases fast evolving third codon positions are not creating more problems to MP than first and second codon positions combined, and that they contain useful phylogenetic signal. This has also been demonstrated in other plant data sets (Savolainen et al., 2002). PHT indicated that no conflict existed between first/second and third codon partitions of all DNA regions, except *rbcL* and *gbss*, and the results of the Monte Carlo simulations indicated that the effect of combining the three codon positions in most sequences had a beneficial impact in reducing error (Fig. 5.4). An increase in error would have been expected if conflict was present between codon partitions, as it is the case for example with *gbss* (Fig. 5.4). The six DNA regions used varied greatly in length (Table 5.1) and in their intrinsic rate of evolution (Fig. 5.3), but most of them created many problems to MP in order to retrieve the model tree (Fig. 5.4). It means that those data have little power to discriminate among the two alternative hypotheses, and therefore it would lessen the confidence associated with the inference.

5.4.4 Long-branch attraction

Increasing the size of the simulated data sets to 5,000 and 10,000 bp resolved the reconstruction problems encountered by MP for *rbcL*, *rpoC2* and ITS2+5.8S, and the percentage of time the correct trees were found with these data sets increased to 90% or higher (Fig. 5.5). This indicates that these three DNA regions were only affected by random error, and that MP would be able to find the correct tree if enough data was sampled. However, *ndhF*, *phyB,* and *gbss* remained problematic when the number of nucleotides was increased. A bias towards the BE-P model tree was found with *ndhF* and *phyB*, but although this bias was reduced when the number of nucleotides was increased for *phyB*, the bias increased with *ndhF* (Fig. 5.5). This indicates that MP is inconsistent with the plastid *ndhF* region, and therefore selected the wrong topology with a higher probability as the number of characters increased. It is clear from Figure 5.3 that, with *ndhF*, terminal taxa in all clades have longer branch length than the internal branches connecting the major clades of the family. This could indicate a rapid radiation of the grasses after the origin of the family (Jacobs et al., 1999), promoted by key innovations such as formation of intercalary meristems or the acquisition of mechanisms for drought tolerance (GPWG, 2001). Such heterogeneity in branch lengths opens the possibility of long-branch attraction in phylogeny reconstructions. Judiciously breaking long branches by adding new taxa has been proposed as means of removing this problem (Hillis, 1996; Kim, 1996; Graybeal, 1998; Rannala et al., 1998), and this approach proved useful with the *ndhF* data sets (Fig. 5.5). However, the number of taxa needed to be high in order to reduce the bias found in *ndhF*. The grass family is one of the largest families in the angiosperms and contains around 10,000 species.

Long-branch attraction is likely to play an important role in our attempts to understand the evolutionary history of the family unless more effort is made to have a better representation of this immense familial diversity. The positions of other problematic taxa (e.g. *Anomochloa*, *Pharus*, *Streptogyna*) and clades (e.g. arundinoids) that are still puzzling grass systematists (GPWG, 2001) could be the result of the same problem and more thorough sampling is necessary.

Our simulations did not bring a clear solution to the central question of whether wheat is closer to rice (BE-P) or closer to maize (PACCAD-P), and given the high rate of error encountered with analyses based on each sequence taken separately, it is doubtful that a single DNA region of short length (< 2,000 bp) could give an accurate answer. Low rates of error were achieved when the different plastid sequences were combined (Fig. 5.4), and given that both ML and MP found the PACCAD-P topology with this data set, there is a slight edge in terms of weight of evidence for this hypothesis. In addition, no morphological synapomorphies supporting the BE-P clade have been identified. In contrast, the standard grass spikelet is present in all members of the PACCAD-P clade, while being variable among other subfamilies, and the loss of the inner stamen whorl could be interpreted as a synapomorphy of the PACCAD-P clade (GPWG, 2001). However, the plastid genomes of wheat, rice, and maize have been completely sequenced, and comparisons of structural features among these three cereal plastomes seem to favour the BE-P hypothesis (Ogihara et al., 2002). It is uncertain how well differences observed among these three species are representative of the rest of the family. In the case of nuclear sequences, high rates of error were detected when the PACCAD-P topology was considered as the model tree (Fig. 5.4). This evidence, along with the low bootstrap percentages obtained for this combined data set (Table 5.2), does not preclude the possibility that the PACCAD-P hypothesis is the true hypothesis underlying the evolution of the nuclear genome. When the number of characters was increased in our simulations, any trace of random error disappeared.

5.5 Conclusion

Studies investigating the deep phylogenetic relationships within grasses, and especially between taxa of immense genetic and economic importance have suffered errors, bias and long-branch attraction in all published work so far; solving this puzzle will probably only be achieved by far more intensive sequencing work and vast increased in taxon sampling.

In the next chapter, the subfamilial relationships within the grasses are investigated using two plastid DNA regions. Following the conclusions of the previous chapters, large phylogenetic trees were reconstructed and the evolutionary hypotheses obtained were used to examine factors that could have influence the diversification of the grass family.

Chapter 6. Grass Evolution: New Insights using Molecules and Fossils

6.1 Introduction

The grasses with 10,000 species and 650 genera are the fifth largest angiosperm family (Clayton and Renvoize, 1986; Mabberley, 1993). Their importance is beyond doubt for they provide the grass-dominated ecosystems, which include tropical and subtropical savannah, temperate grasslands, and steppes, that cover more than a third of the land' s surface (Archibold, 1995) and they play a major role in human sustenance, either as cereal crops or as a source of forage (Raven et al., 1992). The success of the grass family can be explained in part by their adaptability to changeable environments, their ability to resist grazing and coexist with man and by almost endless morphological variations based on an 'all-purpose plant body' (Clayton and Renvoize, 1986; Chapman, 1996).

6.1.1 Grass phylogenetics

Many comprehensive classifications of the grass family have been proposed. The first taxonomic descriptions were not reliant on methods of phylogenetic inference (e.g. Clifford et al., 1969; Clayton and Renvoize, 1986), while more recent hypotheses were based on phenetic (e.g. Watson and Dallwitz, 1992) or cladistic analyses (e.g. Baum, 1987; Kellogg and Campbell, 1987; Doebley et al., 1990; Davis and Soreng, 1993; Liang and Hilu, 1995; Duvall and Morton, 1996; Kellogg, 1998a; Hsiao et al., 1999; GPWG, 2001). Poaceae have been placed consistently with Anarthriaceae, Centrolepidaceae, Ecdeiocoleaceae, Flagellariaceae, Joinvilleaceae and Restionaceae in a group recognised as the order Poales (Dahlgren et al., 1985; Linder, 1987; Doyle et al., 1992; Kellogg and Linder, 1995). However, the sister group of Poaceae has not been clearly established, but some evidence suggests that Joinvilleaceae are their closest relatives (Campbell and Kellogg, 1987; Clark et al., 1995; Soreng and Davis, 1998; GPWG, 2001).

In terms of higher-level groupings within the grass family, a few key findings can be stressed. A small set of taxa including *Anomochloa*, *Guaduella*, Phareae, *Puelia*, and *Streptochaeta* represent a varying number of early-diverging lineages within Poaceae and generally appear as sisters to the remainder of the family (Clark et al., 1995; Duvall and Morton, 1996; Soreng and Davis, 1998; Clark et al., 2000). These taxa were grouped with the woody bamboos and the Olyreae by Clayton and

Renvoize (1986) to form Bambusoideae *s.l.*. Several subfamilies, such as Bambusoideae *s.s.*, Chloridoideae, Panicoideae, and a 'core' Pooideae group, have been recognised as monophyletic by almost all molecular analyses (Barker et al., 1995; Clark et al., 1995; Duvall and Morton, 1996; Mathews and Sharrock, 1996; Soreng and Davis, 1998; Hsiao et al., 1999; GPWG, 2001).

Grass subfamilies, except the ones that are considered part of the early-diverging lineages, have been grouped into two higher clades. A PACCAD clade, comprising Panicoideae, Arundinoideae, Chloridoideae, Centothecoideae, Aristidoideae, and Danthonioideae, has been resolved by most analyses (e.g. Davis and Soreng, 1993; Clark et al., 1995, Duvall and Morton, 1996; GPWG, 2001). Most recent phylogenetic analyses have refuted the monophyly of a broadly defined Arundinoideae (Barker et al., 1995; Clark et al., 1995; Barker et al., 1999; Soreng and Davis, 1998; Hsiao et al., 1999), with some genera being placed far from the core of the group (e.g. *Anisopogon* with Pooideae or *Gynerium* with Panicoideae). However, the bulk of the arundinoids can be associated as a small number of closely related lineages (Hsiao et al., 1998; GPWG, 2001). According to the GPWG (2001), a BEP clade groups the remaining subfamilies (i.e. Bambusoideae, Ehrhartoideae and Pooideae). In this clade, the core of the Pooideae (e.g. Aveneae, Bromeae, Poeae and Triticeae) have been supported by most studies (Soreng et al., 1990; Davis and Soreng, 1993; Soreng and Davis, 1998), but a set of disparate elements traditionally assigned to Arundinoideae and Bambusoideae (e.g. *Anisopogon*, *Diarrhena* and Stipeae; Davis and Soreng, 1993; Clark et al., 1995; Soreng and Davis, 1998) formed a series of close lineages associated with the core Pooideae. Therefore, a large group has emerged around the core Pooideae by including elements not formally included in this subfamily (GPWG, 2001). However, the BEP clade represents one of the major areas of controversy in grass phylogenetics. More particularly, the recent molecular studies have resulted in discrepancies between gene trees concerning the placement of the Pooideae *s.l.* in reference to the PACCAD clade. This problem is precisely the subject of chapter 5 of this thesis, where a thorough investigation of the problem was made. Further discussion of taxonomic groupings can also be found in chapter 4.

6.1.2 Grass evolution

The appearance of grasses happened relatively late in the evolutionary history of angiosperms and the first evidence from fossil deposits includes whole plants with spikelets and inflorescences that are dated from early Eocene (~55 Mya;

Thomasson, 1987; Crepet and Feldman, 1991). These deposits represent the earliest unequivocal macrofossils of grasses, but ambiguous pollen grains that could be associated with grasses has been recorded as far back as the late Cretaceous (~65 Mya; Daghlian, 1981). However, the global expansion of grasses and their increasing relative abundance in earth' s ecosystems did not take place before the early to middle Miocene (~15 Mya; Janis, 1993; Webb et al., 1995; Jacobs et al., 1999; Willis and McElwain, 2002). There is also evidence from plant macrofossils, pollen, and phytoliths (Jacobs et al., 1999) as well as the dentition and skeletal structure of fossil vertebrates (Janis et al., 2000) indicating that all major grass subfamilies had evolved by the early Miocene (~25 Mya).

A number of hypotheses have been proposed to explain this relatively late evolution of the grasses. For instance, the balance of global fauna and the radiation of mammals during the late Cretaceous may have an important role in the expansion of grasslands and increased speciation (Janis, 1993). In particular, coevolution between grasses and hoofed mammals has been hypothesised (Chapman, 1996; MacFadden, 1998; Wing and Boucher, 1998). One of the most compelling arguments suggests that increasing latitude aridity associated with lower temperatures promoted the evolution and expansion of the grasses (Wing and Boucher, 1998). Morphological characteristics of grasses may have been a major selective advantage in conditions of global aridity (Archibold, 1995). Anatomical characteristics may also have been important during this period, and the appearance of the C_4 photosynthetic pathway in grasses could have promoted their expansion as well. C_4 is a particular type of photosynthesis that is present in 18 angiosperm families, with a majority of species found in grass family (Kellogg, 1999). The advantage conferred by C_4 consists of a more efficient CO_2 uptake and reduced water loss under conditions of high temperatures and low precipitation than C_3 plants (Larcher, 1995). C_4 photosynthesis evolved relatively late with the first fossil evidence dating back to the middle Miocene (~16 Ma) and a possible rapid expansion of C_4 plants followed in the late Miocene (~7 Mya; Cerling et al., 1993; MacFadden and Cerling, 1994). This relatively late evolution of the C_4 compared with the C_3 pathway, which was the main photosynthetic pathway of the earliest terrestrial plants (Raven et al., 1992), has been the subject of much debate (Willis and McElwain, 2002). The hot and dry climate typical of the Miocene (Cerling et al., 1997) has often been linked with the appearance of C_4 metabolism, but hot and arid climatic conditions are found many times during early earth history without any clear evidence supporting C_4 evolution before the Miocene (Bocherens et al., 1996; Willis and McElwain, 2002). Other environmental or biological factors must have played a role and one of the main hypotheses is a reduction of atmospheric CO_2 levels during the Miocene, which would have put C_4 plants at a selective advantage (Cerling et al., 1997).

6.1.3 Timing of divergence events

For nearly four decades, the concept of the molecular clock (Zuckerkandl and Pauling, 1965a,b) has been applied to infer divergence dates (Soltis et al., 2002b), but it has become clear that many estimated divergence times are inconsistent with the fossil record (Martin et al., 1993; Heckman et al., 2001; Rodriguez-Trelles et al., 2002). For instance, most molecular-based estimates of the age of the angiosperms greatly exceed the dates inferred from the fossil record (Gaut et al., 1992; Sanderson and Doyle, 2001). Rejection of the standard Poisson process associated with the molecular clock has accumulated thanks to the development of relative-rate tests (Li and Graur, 1991; Gaut et al., 1992; Muse and Weir, 1992; Clegg et al., 1994, Li, 1997). The reason for the rejection of the model may have been that the constant-rate assumption failed. Due to differential evolutionary processes in different lineages, the rate of evolution could still be modelled with a Poisson process, with different rates throughout the tree. However the model may also have been rejected because the Poisson assumption failed. In that case, we can still assume a constant rate of evolution throughout the tree but with an overdispersed stochastic process underlying the evolution of DNA sequences (Gillespie and Langley, 1979). Therefore, new approaches have consisted of building explicit models of overdispersion of the process (Cutler, 2000). Two different ideas have been used to tackle this problem. On one hand, Cutler' s (2000)model assumed that the number of substitutions in lineages was stationary, but, unlike a Poisson process, the variance in the number of substitutions would not necessarily equal to the mean. On the other hand, models have been built with the assumptions that the substitution process was basically Poisson, but that rates varied between lineages (Hasegawa and Kishino, 1989, Lynch and Jarrell, 1993; Uyenoyama, 1995; Sanderson, 1997, 2002; Thorne et al., 1998). Further discussion of timing of divergence dates can be found in the introductory chapter 1. One possible approach for modelling variation in the evolutionary rates between lineages is to place a constraint on the temporal autocorrelation of rates in lineages (Sanderson, 1997). For example, the molecular clock hypothesis assumes a very strong constraint by imposing a constant rate across the tree. However, it is possible to weaken this constraint, but still keep it sufficient to allow estimation of divergence times (Sanderson, 2002a). This is the approach used by penalized likelihood (PL), which is a semiparametric technique. PL takes a parameter-rich model that would over-fit the data and constrains fluctuations in its parameters by a roughness penalty (Green and Silverman, 1994), which forces rates to change smoothly from branch to branch. The parametric component of the method makes it possible to examine a broad spectrum of solutions with different levels of rate smoothing, ranging from highly penalized (i.e. nearly constant rates), to nearly unconstrained rate variability. The introduction of this smoothing parameter avoids over-fitting the data and allowing too much rate variation, which is the characteristic of nonparametric approaches (Sanderson, 2002a). However, deciding the optimal level of smoothing to use is a critical part of the process, and a cross-validation procedure can be used to define its adequate value (Sanderson, 2002a). The idea is to prune terminal branches from the tree one after another, but leaving the immediate ancestral node in place. PL is then performed using predefined levels of smoothing, and the predicted branch length of the pruned terminal is compared with the observed length. The value of smoothing minimising the differences between the predicted and observed branch lengths is considered as optimal (Sanderson, 2002a). The cross-validation can be used repeatedly to narrow the range around the optimal value.

Despite theoretical advances, divergence times estimated from molecular data have to be taken with caution, and several factors can influence the age estimates (Sanderson and Doyle, 2001). A lower bound on errors in age estimates is imposed by the DNA substitutional process itself (Hillis et al., 1996), but this problem can be reduced by using appropriate models of DNA evolution (Sanderson and Doyle, 2001). However, errors in the estimation of the underlying evolutionary hypothesis, as well as in the assignment of fossil dates to nodes in the tree, cannot be dismissed and are likely to have an impact on the divergence times estimated (Sanderson and Doyle, 2001).

6.1.4 DNA sequences used in grass phylogenetics

Several DNA regions from the different genomes have been used to infer the evolutionary history of the grasses. Chapter 5 examined the phylogenetic information contained in six DNA regions from the plastid and nuclear genomes, but additional sequences such as the plastid *rps4* (Nadot et al., 1994), *matK* (Liang and Hilu, 1995; Hilu et al., 1999) and *rps16* (Zhang, 2000) or the nuclear *Adh* (Gaut et al., 1999), have also been used in recent molecular studies. In this chapter, two plastid DNA sequences were used. The transfer RNA region sequenced here, called *trnLF*, represents a continuous section of DNA including the partial coding sequence of *trnL* 3' and 5' exons (UAA anticodon), the *trnLF* intergenic spacer and the partial coding sequence of the *trnF* exon (GAA anticodon). This DNA region has been used in studies investigating higher-level relationships within the grasses (Doust and Kellogg, 2002; Hodkinson et al., 2002b), but has not been used at a wider taxonomic level. The *rbcL* region sequenced here represents the large subunit of ribulose-1,5-bisphosphate carboxylase, an enzyme involved in carbon fixation. It is one of the DNA regions most widely used in angiosperm and land plant phylogenetics (e.g. Chase et al., 1993; Källersjö et al., 1998; Savolainen et al., 2000), and has been used previously on a smaller taxonomic sample of the grass family (Duvall and Morton, 1996).

6.1.5 Aims

Phylogenetic analyses of individual and/or combined molecular data sets have converged on a set of well-supported relationships with the grasses that led to the first family-wide subfamilial classification based on explicit phylogenetic hypotheses (GPWG, 2001). However, the taxonomic representation in these analyses remained low, and the aims of this chapter were to increase the sampling of taxa in order to have a more comprehensive understanding of the relationships within the family and to provide evidence based on a the *trnLF* region and an extended taxa sampling for the *rbcL* region. Three large DNA matrices, represented by these two plastid DNA regions, analysed separately, and in combination, were used to infer the phylogenetic relationships within the grass family. These evolutionary hypotheses were then used to date the divergence events within the family, and to examine the effects of levels of past CO_2 concentration on the origin of the C_4 photosynthetic pathway.

6.2 Material and Methods

6.2.1 Specimens and DNA extraction

Specimens were collected from the living collections at the Royal Botanic Gardens, Kew, UK and during different field trips by Trevor R. Hodkinson. Voucher specimens, Kew and Trinity College DNA bank numbers, as well as GeneBank accession numbers when available, of each accession are listed in Appendix 6.1.

DNA was extracted from 0.5-1.0 g of silica gel (Sigma) dried leaf material using a modified 2X CTAB procedure of Doyle and Doyle (1987), precipitated using 100% ethanol or isopropanol for at least 48 hours at –20°C, pelleted and washed with 70% ethanol and purified via cesium chloride/ethidium bromide (1.55 g/ml) gradient centrifugation with subsequent dialysis to remove salts. Some DNAs were also purified using Concert PCR purification columns (Gibco BRL). Ethidium bromide was extracted with H_2O-saturated butanol. DNA was then stored in TE buffer (10 mM Tris-HCl; 1 mM EDTA; pH 8.0) at –80°C until use.

6.2.2 DNA sequencing

The spacer and intron of the plastid *trnLF* region were amplified as one piece using the c forward (5'-CGAAATCGGTAGACGCTACG-3') and f reverse (5' ATTTGAACTGGTGACACGAG-3') primers described by Taberlet et al. (1991.) The thermal cycling comprised 30 cycles, each with 1 min. denaturation at 97°C, 1 min. annealing at 51°C. and an extension of 3 min. at 72°C. A final extension of 7 min. at 72°C was also included. Amplified, double-stranded DNA fragments were purified using Concert PCR purification columns (Gibco BRL) and sequenced using *Taq* Dye-Deoxy Terminator Cycle Sequencing Kits of Applied Biosystems on an Applied Biosystems 310, 373, or 377 automated DNA sequencer, all according to the manufacturer's protocols and with the same primers as the initial amplification. Sequence editing and assembly of the complementary strands used Sequence Navigator and AutoAssembler programs (Applied Biosystems) or the Staden package (Medical Research Council, Laboratory of Molecular Biology). Each position was individually inspected to be sure that both strands agreed. The 189 DNA sequences for *trnLF* were a collaborative effort between Grainne NiChongaile, Trevor R. Hodkinson and myself.

The *rbcL* data set was sequenced at the Royal Botanic Gardens, Kew as part of a wider project aiming at understanding the relationships within the monocots (V. Savolainen; pers. comm.). These sequences were extracted and sequenced by Dr. Michelle van der Bank. Moreover, several sequences for *rbcL* and *trnLF* regions were downloaded from GeneBank to complement the sampling of the different grass subfamilies (Appendix 6.1).

6.2.3 Data analyses

For *trnLF*, 189 sequences were aligned using CLUSTAL X (Thompson et al., 1997) with subsequent manual correction following the guidelines of Kelchner (2000). Gaps smaller than 10 bp were coded as missing data, while larger ones were excluded from the analysis. The 177 sequences for *rbcL* were aligned by hand. A combined matrix was created with taxa in common between the *trnLF* and *rbcL* matrices. On the 123 taxa present in the combined matrix, seven key taxa for *trnLF* and 22 for *rbcL* were included even though they were lacking the respective DNA region (i.e. missing characters were added to these taxa for one of the DNA regions). Within genera, some species had been sequenced for one DNA region but not for the other. However, a different species from the same genus was available and these were concatenated into one entry in the combined matrix when they formed a monophyletic genus in the separate analyses. The different taxon sampling between the two matrices is due to the fact that *trnLF* and *rbcL* matrices were produced independently, and that the collaboration took place only after the taxon sampling was done.

The three resulting matrices were analysed by MP using heuristic search options as implemented in PAUP*4b. Searches included 1000 replicates of random addition sequence (saving no more than 100 trees per replicate to reduce time spent swapping large islands of trees) with TBR branch-swapping on multiple trees. Internal support was assessed using 1,000 bootstrap replicates (Felsenstein, 1985a). Searches included 10 replicates of random addition sequence (saving no more than 5 trees per replicate to reduce time spent swapping on large numbers of trees) with TBR branch-swapping. The suitable model of DNA substitution for each DNA sequence was determined with ModelTest3.6. The HKY85+Γ fitted the *rbcL* data best, while the GTR+Γ+I was selected for *trnLF* and the combined matrix. The parameters of the different models are presented in Appendix 6.2. ML branch lengths and parameters of the models of DNA substitution were estimated on all MP trees found, and the tree having the highest likelihood was selected for the subsequent dating procedures. Bayesian analyses were also performed on the three matrices using the models described above using MrBayes2.1. The MCMC algorithm was run during 500,000 generations on four chains. Trees were sampled every 100 generations and the burn in period was determined for each matrix by inspecting how many generations were required for the likelihood function to stabilise. The Bayesian analyses were done on the IBM NetFinity 32 node cluster available at University of Dublin, Trinity College.

6.2.4 Estimation of age of divergence

Four grass fossils were chosen to calibrate the phylogenetic trees. The earliest evidence for grasses in the fossil record have been dated between 55 to 65 Mya and are represented by spikelets and inflorescence fragments including pollen (Crepet and Feldman, 1991; Jakobs et al., 1999). Remains of a fossil grass assigned to the genus *Stipa* have been dated at 33 Mya (Leopold et al., 1992; Jakobs et al., 1999), while two fossils representing the genera *Distichlis* and *Cleistochloa* were dated at 12.5 Mya (Dugas and Retallack, 1993; Jakobs et al., 1999). The ages of clades estimated using these calibration points were lower bounds on the divergence times estimated conservatively for the crown group by the first appearance of fossils clearly referable to one of the constituent lineages based on morphological synapomorphies (Soltis et al., 2002b).

The method of PL (Sanderson, 2002a) as implemented in r8s version 1.5 (Sanderson, 2002b) was used to estimate absolute dates from the branch lengths obtained from the three matrices described above and calibrated with the four grass fossils. The smoothing parameter necessary to perform PL was estimated using a cross-validation procedure (Sanderson, 2002a). To estimate the confidence intervals associated with each date, 100 bootstrap trees were obtained by resampling the three matrices using ML. Each bootstrap matrix was then optimized on the observed trees obtained for *trnLF*, *rbcL* and the combined data to obtain new age estimates by PL using r8s1.5 as described above. The parameters of the different models were set at the value found in the original matrices and were not estimated on each bootstrap replicate to save considerable computational time.

6.2.5 C_4 grasses and past CO_2 levels

Data containing the type of photosynthetic pathway present in each grass taxon used in this chapter was taken from previous morphological descriptions of grass genera (Clayton and Renvoize, 1986; Watson and Dallwitz, 1992). Each taxon was coded as C_3 or C_4 and the character was mapped on the calibrated trees to estimate the date of appearance of the C_4 photosynthetic pathway.

Values of CO_2 concentrations between 5 and 25 Mya have been estimated from marine photosynthetic carbon fixation (Pagani et al., 1999), while values between 55 and 65 Mya have been estimated from stomatal indices (Royer et al., 2001). A model estimating the CO_2 level over the complete period was also considered (Berner and Kothavala, 2001) to bridge the gap between the observed values from the two other studies. The data used by these publications was obtained from the different authors and a statistical analysis for structural change (Maddala and Kim; 1998) in the two time series was performed to find any significant change in the curve of CO_2 levels using the software R version 1.4.0[10]. This analysis works by calculating F statistics for all potential change points along the time series. For each point, two separate regressions are performed for the two subsamples defined before and after the point are fitted, and an F statistic is obtained and compared to an exact F distribution. There are two methods to aggregate the series of F statistics into a test statistic. This idea is to reject the null hypothesis of no structural change when the F statistic gets larger than the maximal or mean F statistic (Andrews, 1993; Andrews and Ploberger, 1994; Hansen, 1997). The asymptotic critical values of the average F test have been tabulated (Andrews, 1993) and this test was chosen because it was easier to implement.

6.3 Results

6.3.1 Phylogenetic analyses

The lengths of the aligned *trnLF*, *rbcL*, and combined data sets were 1794, 1500 and 3294 bp, with 480, 438 and 856 variable sites and 140, 324 and 541 of these were potentially parsimony-informative, respectively.

Figure 6.1 shows the tree obtained from the Bayesian analysis of the *trnLF* data set. In this tree, the PACCAD clade formed a monophyletic group (posterior probability of 1; Fig. 6.1B). Within this clade, Chloridoideae was monophyletic, and was sister to Danthonioideae (posterior probability of 1 for both; Fig. 6.1B). Panicoideae were inferred as a monophyletic group containing two subclades, Andropogoneae and Paniceae (posterior probability of 1, 1 and 0.82 respectively; Fig. 6.1B), and with Centothecoideae as their sister group (posterior probability of 0.99; Fig. 6.1B). The sister group of the PACCAD clade was Ehrhartoideae, but with a low posterior probability of 0.51 (Fig. 6.1A). Pooideae were monophyletic (posterior probability of 1; Fig. 6.1C) and were sister to the PACCAD clade and Ehrhartoideae but with a posterior probability of only 0.69 (Fig. 6.1A). Pooideae can be divided into two major clades, one containing Triticeae, and the other Poeae and Aveneae (posterior probability of 1 for both; Fig. 6.1C), with additional sister groups

[10] http://www.r-project.org

such as Meliceae and Stipeae. *Lygeum* and *Milium* were sister to the rest of the Pooideae. Bambusoideae *s.s.* were split in two monophyletic groups containing on one hand the tropical woody bamboos and the herbaceous bamboos as sister to the PACCAD clade, Pooideae and Ehrhartoideae (posterior probability of 0.96; Fig. 6.1D), and on the other hand the temperate woody bamboos as a more basal group (posterior probability of 1; Fig. 6.1E). Therefore, the Bambusoideae *s.s.* were not monophyletic in our analyses. Finally, *Pharus* and *Streptochaeta* formed a monophyletic group (posterior probability of 0.68; Fig. 6.1A) and were basal to the rest of the Poaceae. *Joinvillea* was found as the sister taxa to Poaceae with high confidence (posterior probability of 1; Fig. 6.1A).

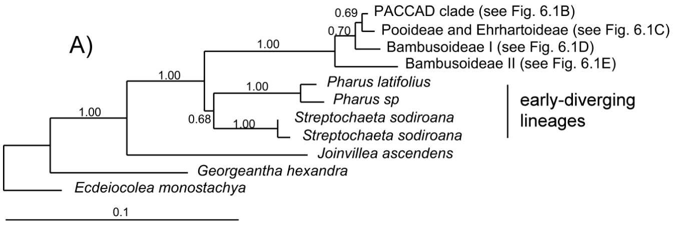


Figure 6.1A

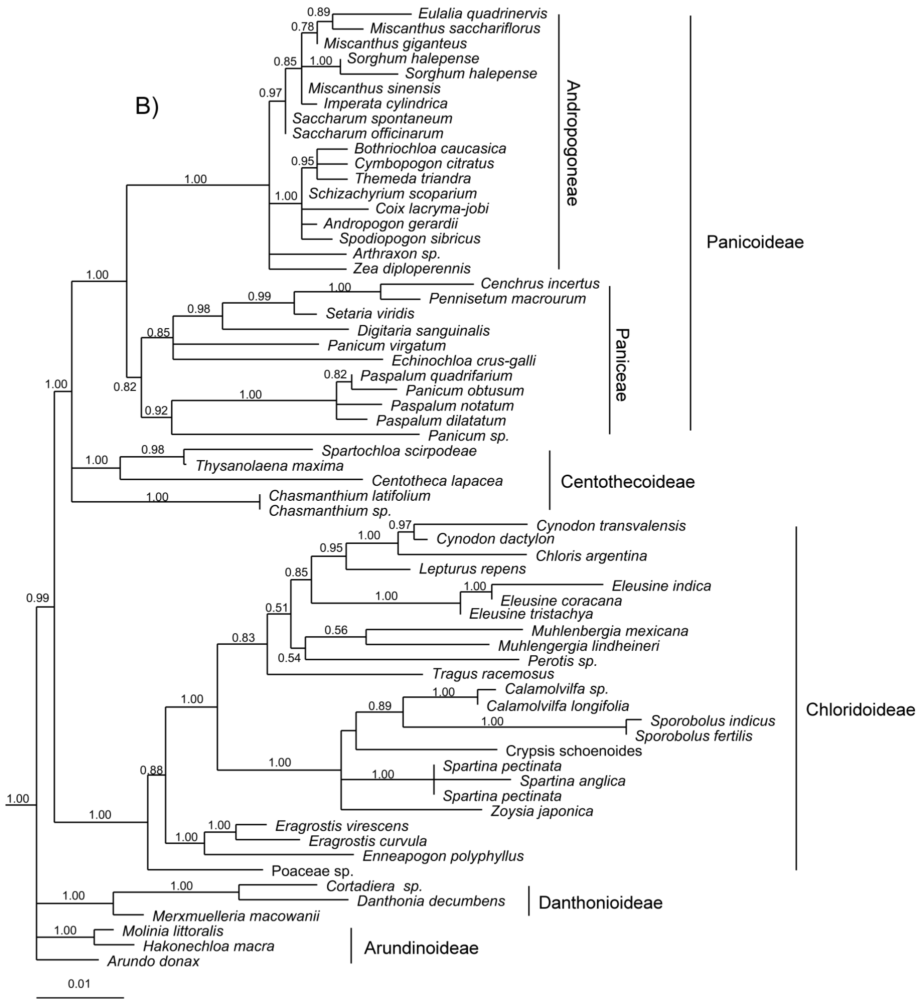

B)
0.89
Eulalia quadrinervis
0.78
Miscanthus sacchariflorus
Miscanthus giganteus
0.85
1.00
Sorghum halepense
Sorghum halepense
0.97
Miscanthus sinensis
Imperata cylindrica
Saccharum spontaneum
Saccharum officinarum
Bothriochloa caucasica
0.95
Cymbopogon citratus
Themeda triandra
1.00
1.00
Schizachyrium scoparium
Coix lacryma-jobi
Andropogon gerardii
Spodiopogon sibricus
Arthraxon sp.
Zea diploperennis
Andropogoneae
Panicoideae
1.00
1.00
Cenchrus incertus
0.99
Pennisetum macrourum
0.98
Setaria viridis
0.85
Digitaria sanguinalis
Panicum virgatum
Echinochloa crus-galli
1.00
0.82
0.82
Paspalum quadrifarium
1.00
Panicum obtusum
0.92
Paspalum notatum
Paspalum dilatatum
Panicum sp.
Paniceae
0.98
Spartochloa scirpodeae
1.00
Thysanolaena maxima
Centotheca lapacea
Centothecoideae
1.00
Chasmanthium latifolium
Chasmanthium sp.
0.97
Cynodon transvalensis
1.00
Cynodon dactylon
0.95
Chloris argentina
0.85
Lepturus repens
1.00
Eleusine indica
1.00
Eleusine coracana
0.51
Eleusine tristachya
0.83
0.56
Muhlenbergia mexicana
Muhlengergia lindheineri
0.54
Perotis sp.
0.99
Tragus racemosus
1.00
1.00
Calamolvilfa sp.
0.89
Calamolvilfa longifolia
1.00
Sporobolus indicus
Chloridoideae
Sporobolus fertilis
Crypsis schoenoides
0.88
1.00
1.00
Spartina pectinata
Spartina anglica
Spartina pectinata
Zoysia japonica
1.00
1.00
1.00
Eragrostis virescens
Eragrostis curvula
Enneapogon polyphyllus
Poaceae sp.
1.00
Cortadiera sp.
1.00
Danthonia decumbens
Danthonioideae
Merxmuelleria macowanii
1.00
Molinia littoralis
Hakonechloa macra
Arundinoideae
Arundo donax
0.01

Figure 6.1B

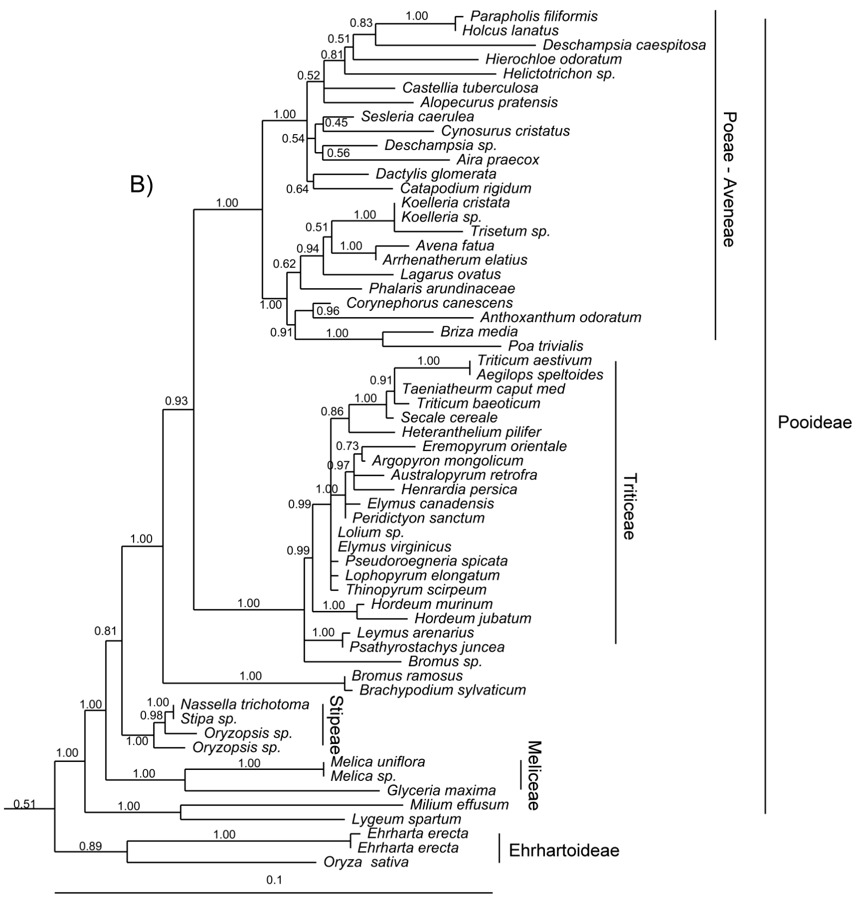


Figure 6.1C

Figure 6.1 - Phylogenetic tree with branch lengths representing expected number of substitution per site obtained from Bayesian analysis based on the trnLF *data set. The posterior probability for each clade are shown on each branch. The tree is separated into five subtrees describing the major clades. A) basal relationships within Poaceae; B) PACCAD clade; C) Pooideae; D) tropical woody bamboos and Olyreae; E) temperate woody bamboos.*

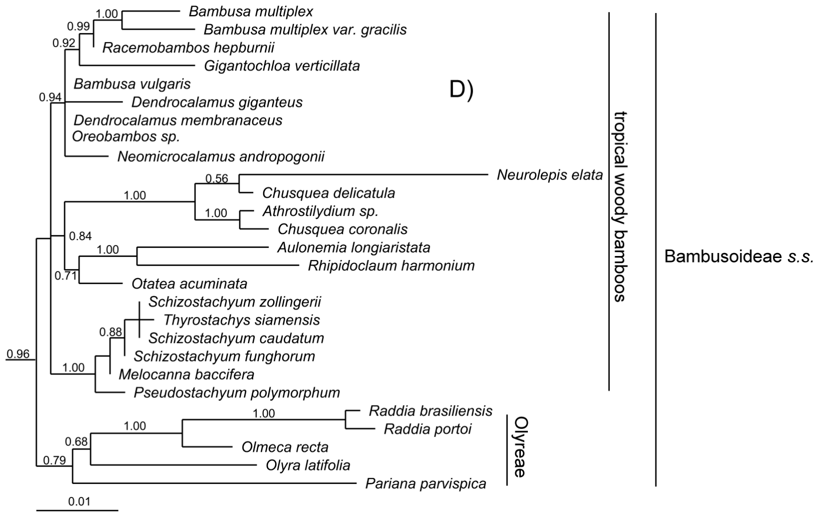


Figure 6.1D

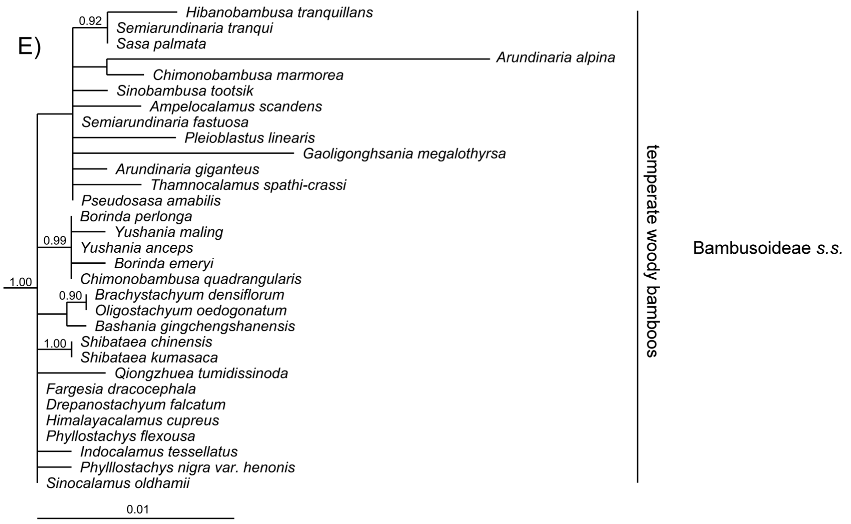


Figure 6.1E

One of 1720 equally most parsimonious trees for *trnLF* is shown in Figure 6.2. It has 4211 steps, with a consistency index of 0.39 and a retention index of 0.71. The results of the MP analysis were different at the subfamilial level, but relationships within each subfamily were similar to the Bayesian analysis (Fig. 6.2). The support obtained by the bootstrap was also lower than the posterior probability in most cases. In the MP analysis, Pooideae were sister to Ehrhartoideae, and embedded in the BEP clade with Bambusoideae *s.s.*, although these relationships were not supported in the bootstrap analysis (Fig. 6.2A). The PACCAD clade was monophyletic (96% of bootstrap; Fig. 6.2B), and contained the monophyletic Chloridoideae and a larger monophyletic clade represented by Panicoideae and Centothecoideae (89 and 61% of bootstrap respectively; Fig. 6.2B). The Panicoideae was not supported as a monophyletic group and the position of Centothecoideae in comparison to Panicoideae was ambiguous (Fig. 6.2).

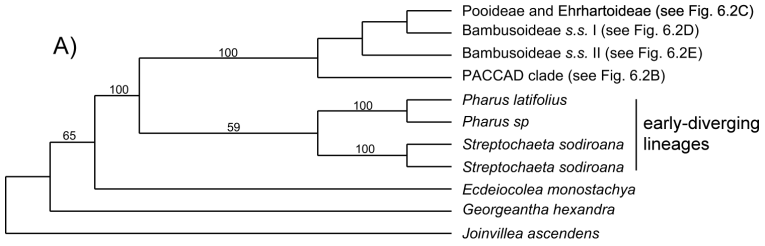

A)
Pooideae and Ehrhartoideae (see Fig. 6.2C)
Bambusoideae *s.s.* I (see Fig. 6.2D)
Bambusoideae *s.s.* II (see Fig. 6.2E)
PACCAD clade (see Fig. 6.2B)
Pharus latifolius
Pharus sp
Streptochaeta sodiroana
Streptochaeta sodiroana
Ecdeiocolea monostachya
Georgeantha hexandra
Joinvillea ascendens
early-diverging lineages
100
100
59
100
100
65

Figure 6.2A

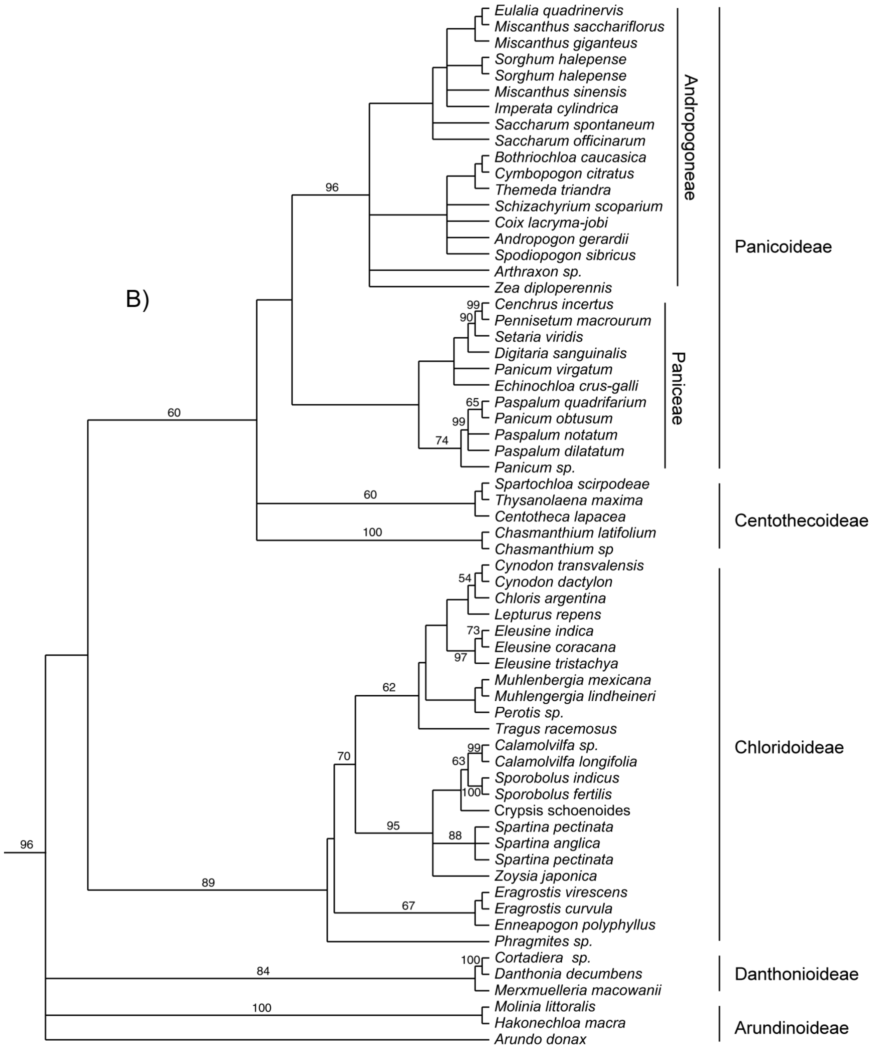

B)
Eulalia quadrinervis
Miscanthus sacchariflorus
Miscanthus giganteus
Sorghum halepense
Sorghum halepense
Miscanthus sinensis
Imperata cylindrica
Saccharum spontaneum
Saccharum officinarum
Bothriochloa caucasica
Cymbopogon citratus
Themeda triandra
Schizachyrium scoparium
Coix lacryma-jobi
Andropogon gerardii
Spodiopogon sibricus
Arthraxon sp.
Zea diploperennis
Cenchrus incertus
Pennisetum macrourum
Setaria viridis
Digitaria sanguinalis
Panicum virgatum
Echinochloa crus-galli
Paspalum quadrifarium
Panicum obtusum
Paspalum notatum
Paspalum dilatatum
Panicum sp.
Spartochloa scirpodeae
Thysanolaena maxima
Centotheca lapacea
Chasmanthium latifolium
Chasmanthium sp
Cynodon transvalensis
Cynodon dactylon
Chloris argentina
Lepturus repens
Eleusine indica
Eleusine coracana
Eleusine tristachya
Muhlenbergia mexicana
Muhlengergia lindheineri
Perotis sp.
Tragus racemosus
Calamolvilfa sp.
Calamolvilfa longifolia
Sporobolus indicus
Sporobolus fertilis
Crypsis schoenoides
Spartina pectinata
Spartina anglica
Spartina pectinata
Zoysia japonica
Eragrostis virescens
Eragrostis curvula
Enneapogon polyphyllus
Phragmites sp.
Cortadiera sp.
Danthonia decumbens
Merxmuelleria macowanii
Molinia littoralis
Hakonechloa macra
Arundo donax
Andropogoneae
Paniceae
Panicoideae
Centothecoideae
Chloridoideae
Danthonioideae
Arundinoideae
96
60
99
90
65
99
74
60
100
54
73
97
62
99
63
100
70
95
88
96
89
67
100
84
100

Figure 6.2B

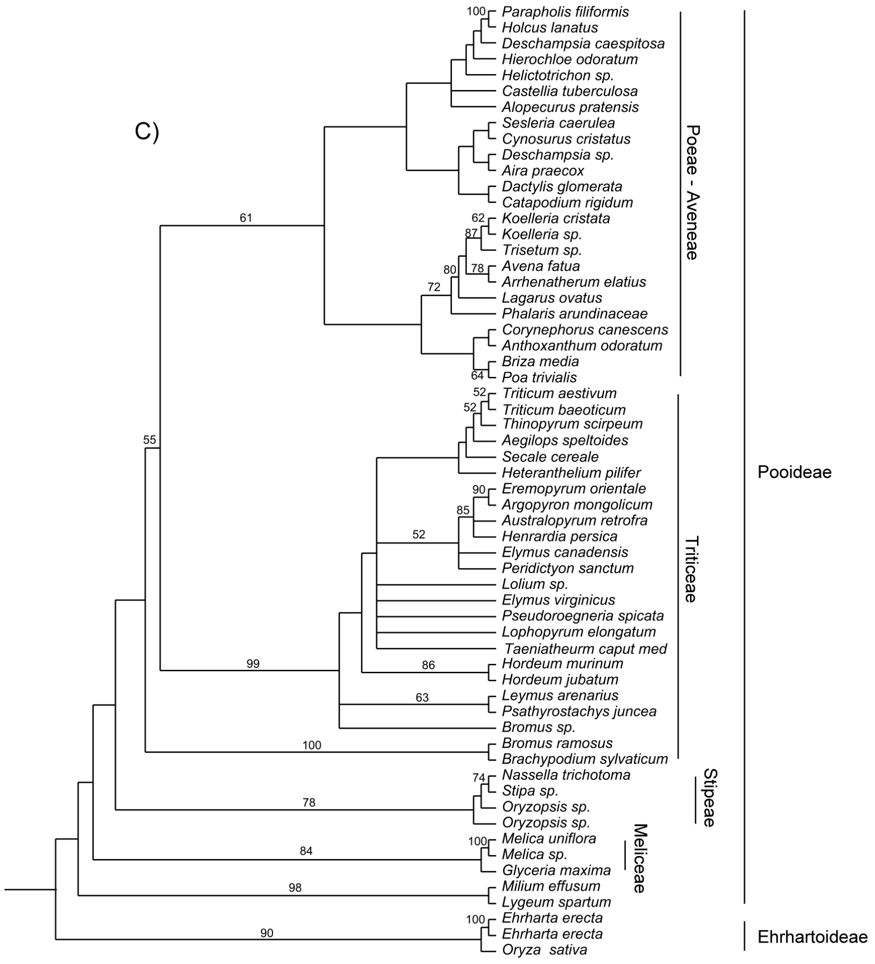


Figure 6.2C

Figure 6.2 - One of the equally most parsimonious trees obtained from MP analysis based on the trnLF *data set. The bootstrap support above 50% is shown on each branch. The tree is separated into five subtrees describing the major clades. A) basal relationships within Poaceae; B) PACCAD clade; C) Pooideae; D) tropical woody bamboos and Olyreae; E) temperate woody bamboos.*

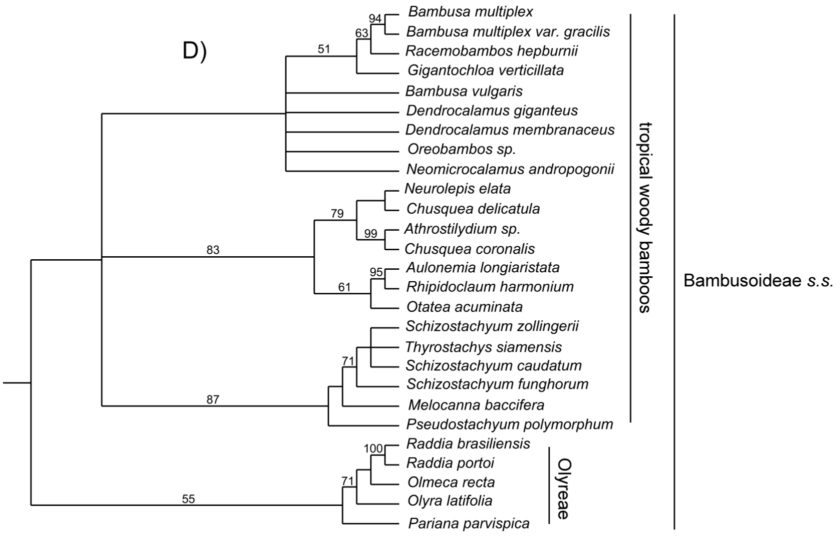


Figure 6.2D

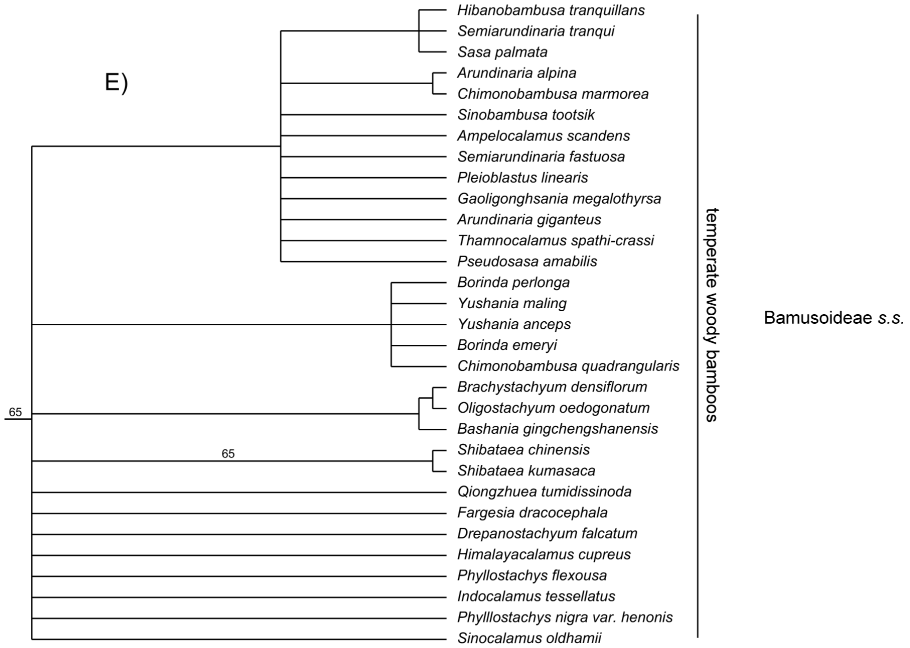


Figure 6.2 E

The MP and Bayesian analyses performed on the *rbcL* data sets were almost identical and only the Bayesian tree is shown with both bootstrap support from the MP analysis and posterior probabilities from the Bayesian analysis (Fig. 6.3). Similarly to the analyses presented in chapter 5, the *rbcL* data sets did not support a BEP clade and Pooideae were placed as sister to the PACCAD clade with high posterior probability but low bootstrap (0.89 vs. 52% respectively; Fig. 6.3A). *Phaenosperma*, a member of Bambusoideae *s.l.*, was basal to all Pooideae (posterior probability of 1 but no bootstrap support; Fig. 6.3A). Within the PACCAD clade, Chloridoideae and Panicoideae were monophyletic (posterior probability of 1 and 0.97; bootstrap support of 84 and 88%; Fig. 6.3.B), while Aristidoideae, Danthonioideae and Arundinoideae were associated in a larger clade with Chloridoideae (posterior probability 0.86 and no bootstrap support; Fig. 6.3B), with the exception of arundinoid *Gynerium* that was sister to all Panicoideae with a posterior probability of 0.97 (Fig. 6.3B). Ehrhartoideae was sister to this PACCAD and Pooideae clade with a posterior probability of 0.95, but no bootstrap support (Fig. 6.3A). Bambusoideae *s.s.* formed a monophyletic group with three clades containing the tropical and temperate woody bamboos and the herbaceous bamboos. However, the relationships between these three groups were not supported (Fig. 6.3C). A clade comprising *Guaduella* and *Puelia*, as well as *Pharus* and a clade comprising *Anomochloa* and *Streptochaeta* formed an increasing set of highly supported inclusive relationships with the remaining core of the Poaceae (Fig. 6.3A). Finally, the sister groups of Poaceae were formed by *Joinvillea, Ecdeiocolea* and *Georgeantha* (Fig. 6.3A).

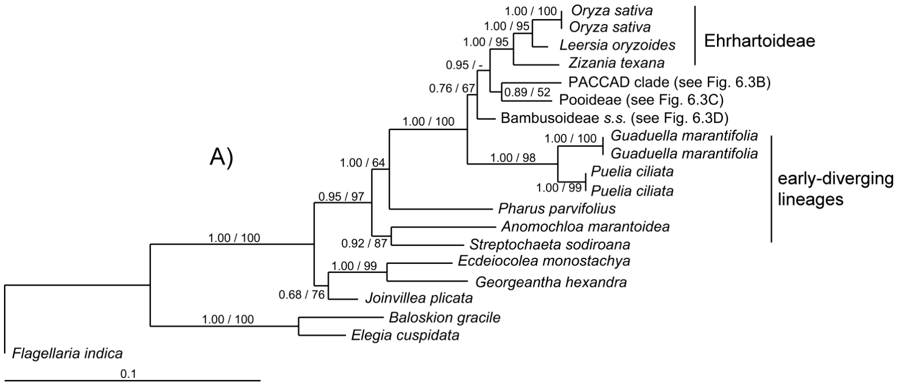


Figure 6.3A

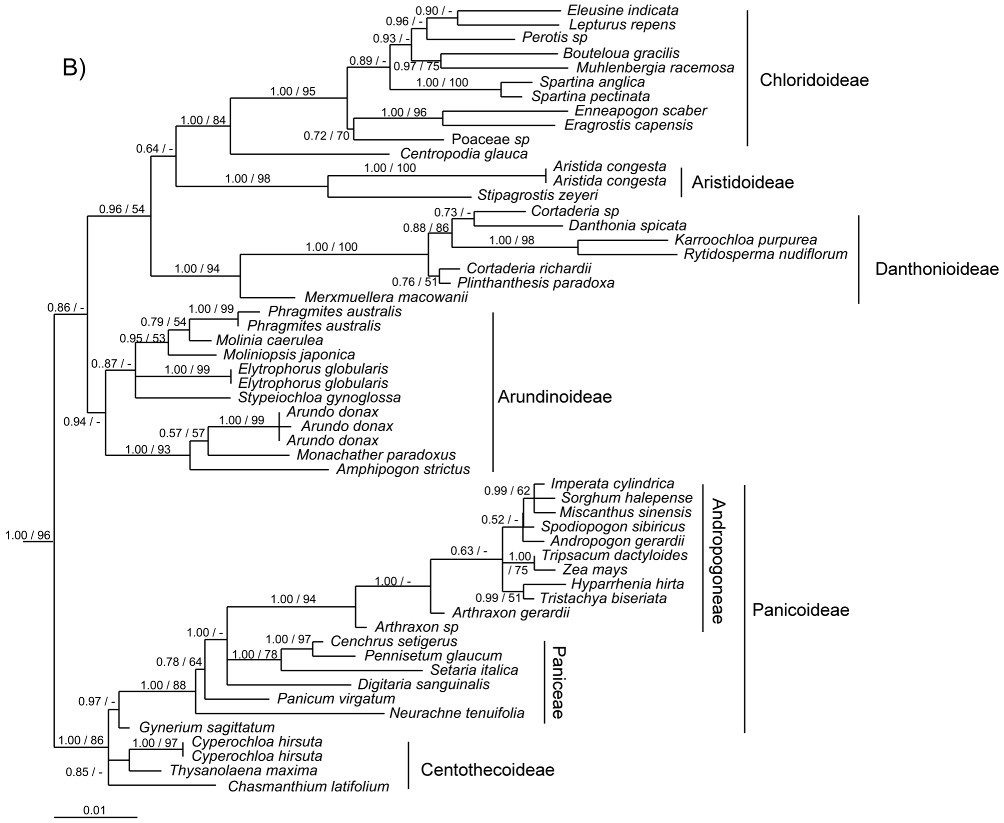
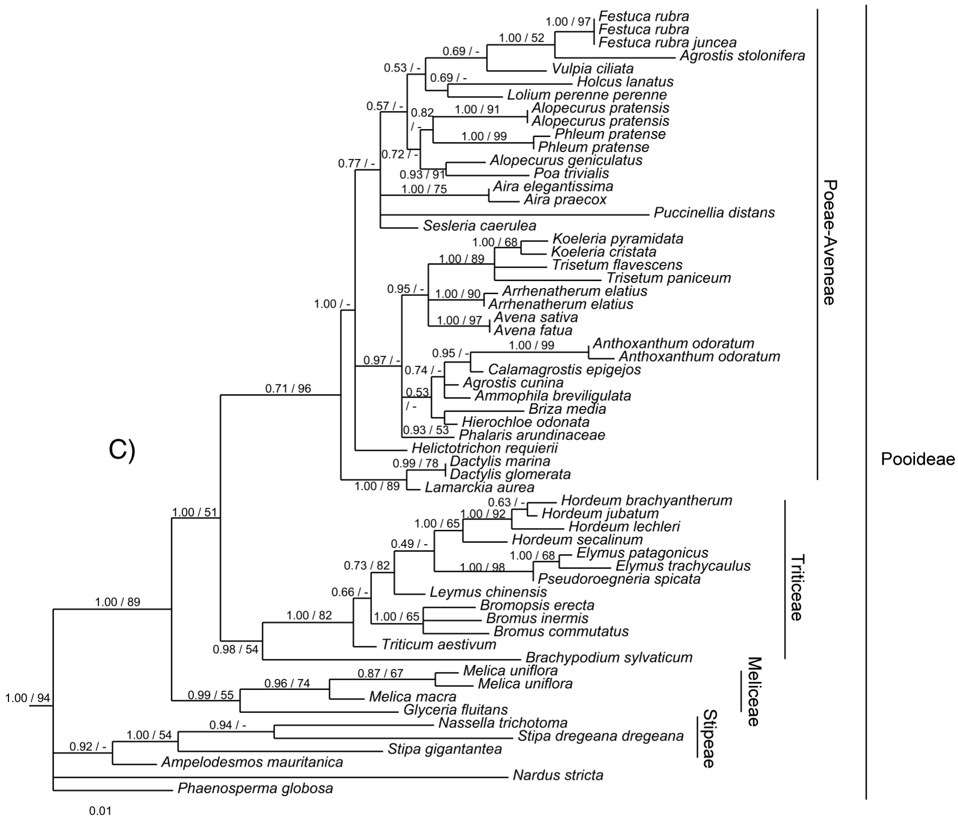


Figure 6.3B (above) and 6.3C (below)



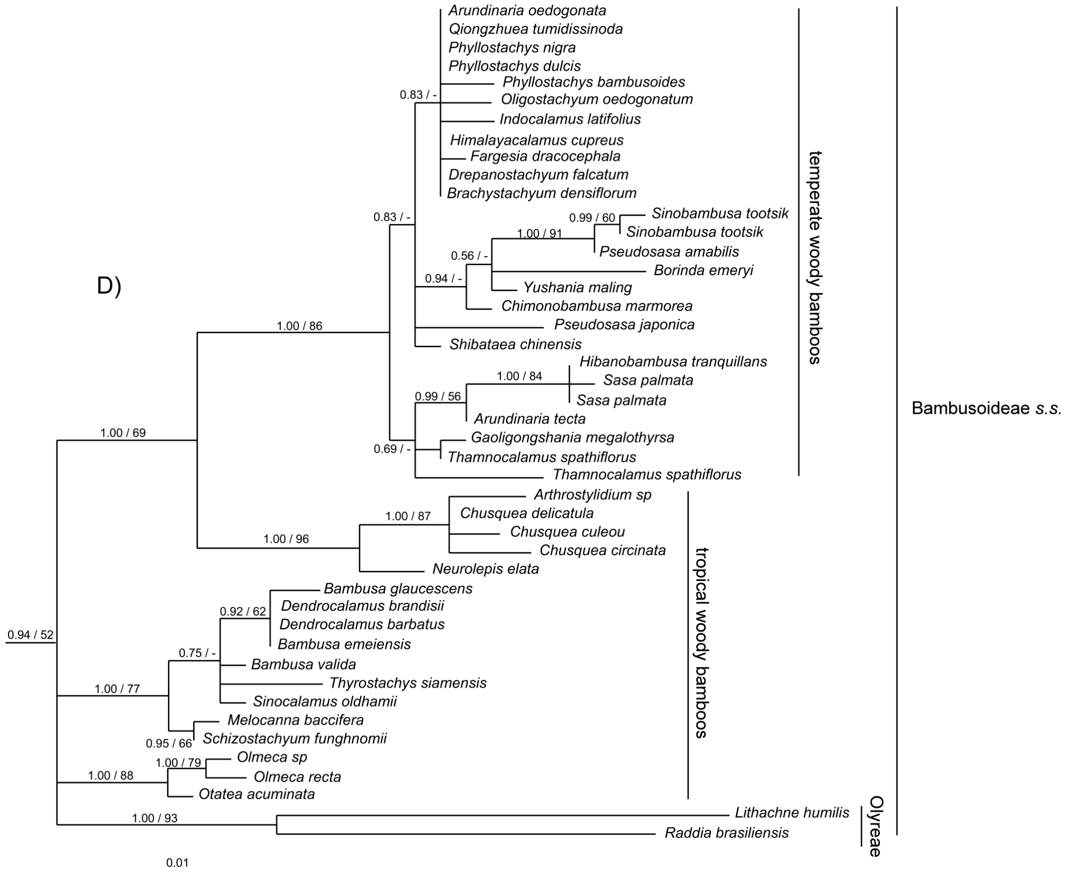


Figure 6.3D

Figure 6.3 - Phylogenetic tree with branch lengths representing expected number of substitution per site obtained from Bayesian analysis based on the rbcL *data set. The posterior probability for each clade are shown on each branch as well as the bootstrap percentages when above 50%. The tree is separated into four subtrees describing the major clades. A) basal relationships within Poaceae; B) PACCAD clade; C) Pooideae; D) Bambusoideae* s.s.

The combined data set resulted in different topologies for the MP or Bayesian analyses, but here again, the relationships within the major groups were similar (Fig. 6.4 and 6.5). With the Bayesian analysis, Pooideae were sister to the PACCAD clade (posterior probability of 1; Fig. 6.4A). Within the PACCAD clade, Panicoideae formed a highly supported monophyletic group (posterior probability of 1; Fig. 6.4B) sister to Centothecoideae; *Gynerium* was again sister to all Panicoideae (Fig. 6.4B). Chloridoideae formed a monophyletic group, but only with a posterior probability of 0.49 (Fig. 6.4B). Danthonioideae and Arundinoideae *s.l.* were associated in a larger clade with Chloridoideae (posterior probability 0.87; Fig. 6.4B). The sister clade of the PACCAD clade and Pooideae was Ehrhartoideae although the posterior probability was very low (0.35; Fig. 6.4A). Bambusoideae *s.s.* formed a monophyletic group containing two clades, the temperate woody bamboos, and the tropical woody bamboos and the herbaceous bamboos (posterior probability of 0.82 and 0.63 respectively; Fig. 6.4D). Sister to Bambusoideae *s.s.* was a clade containing *Guaduella* and *Puelia*, although the posterior probability was again low (0.44; Fig. 6.4A). The remaining early-diverging lineages of grasses were sister to the core Poaceae with first *Pharus* followed by *Anomochloa* and *Streptochaeta* that were grouped together (Fig. 6.4A). Finally, *Ecdeiocolea* and *Georgeantha* (both Ecdeiocoleaceae) were considered as the sister group of Poaceae (Fig. 6.4A).

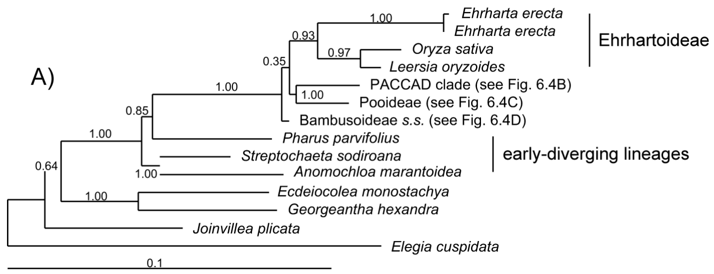


Figure 6.4A

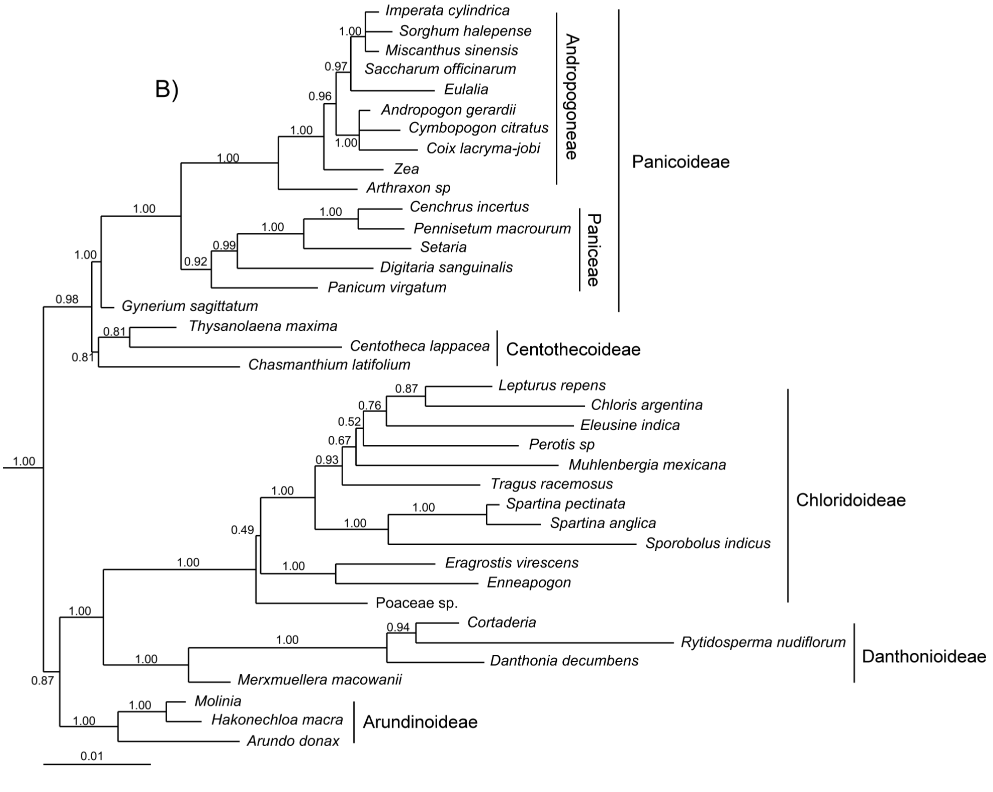


Figure 6.4B

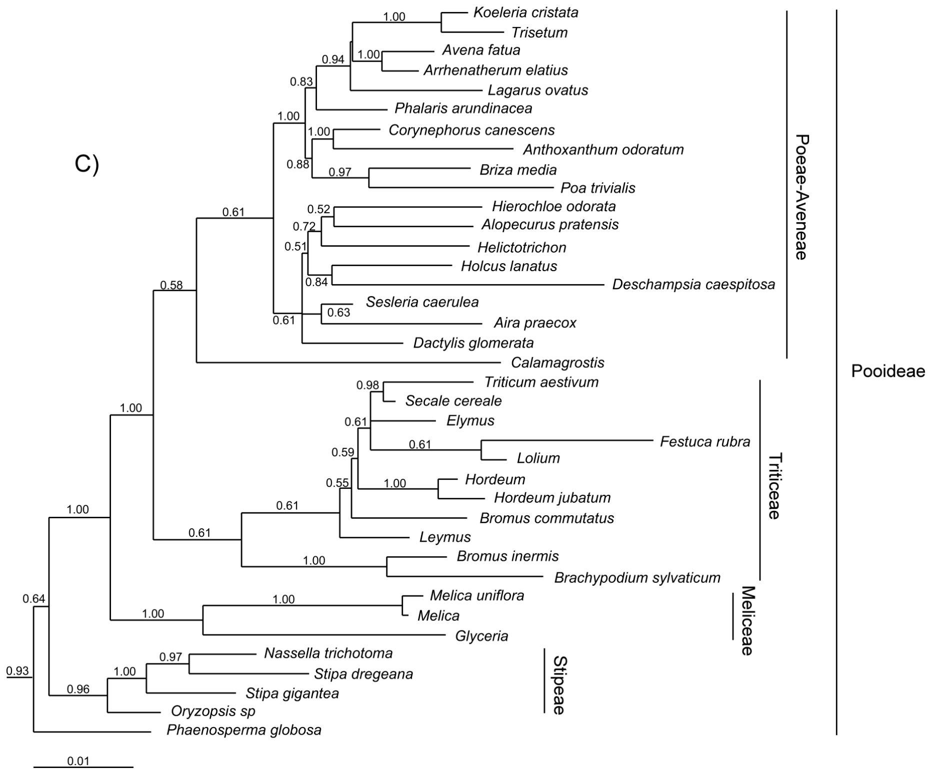


Figure 6.4C

Figure 6.4 - Phylogenetic tree with branch lengths representing expected number of substitution per site obtained from Bayesian analysis based on the combined data set. The posterior probability for each clade are shown on each branch. The tree is separated into four subtrees describing the major clades. A) basal relationships within Poaceae; B) PACCAD clade; C) Pooideae; D) Bambusoideae s.s*.*

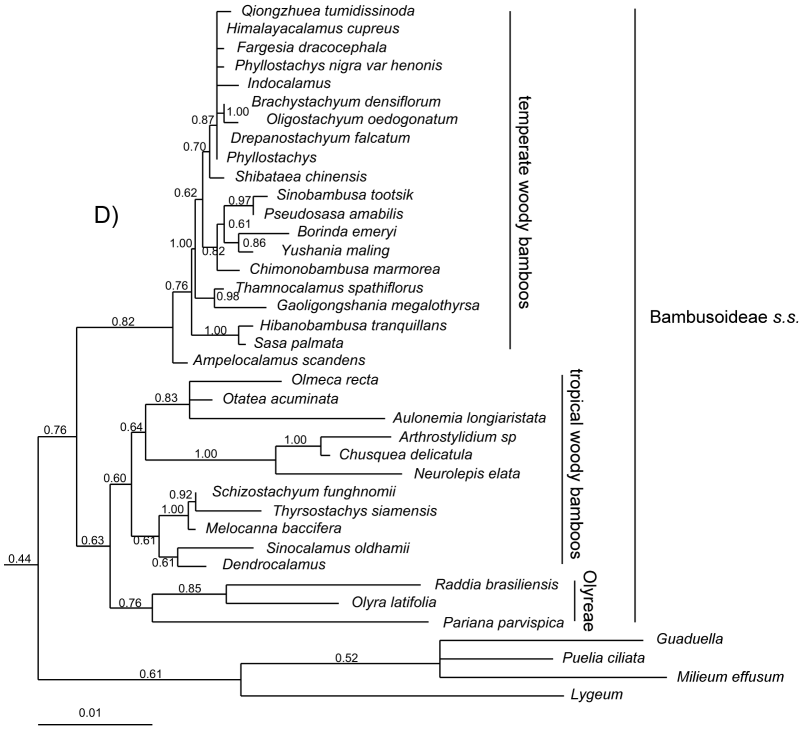


Figure 6.4D

One of 4600 equally most parsimonious trees for combined matrix is shown in Figure 6.5. It has 2532 steps, with a consistency index of 0.47 and a retention index of 0.75. The MP analysis differed in two main points. Firstly, Pooideae were not grouped with the PACCAD clade but within the BEP clade, as sister to Bambusoideae (Fig. 6.5A). Secondly, *Guaduella* and *Puelia* were grouped together as sister to the core Poaceae (Fig. 6.5A). However, and in contrast to the posterior probabilities, the bootstrap analysis showed almost no support throughout the tree except for Panicoideae (Fig. 6.5).

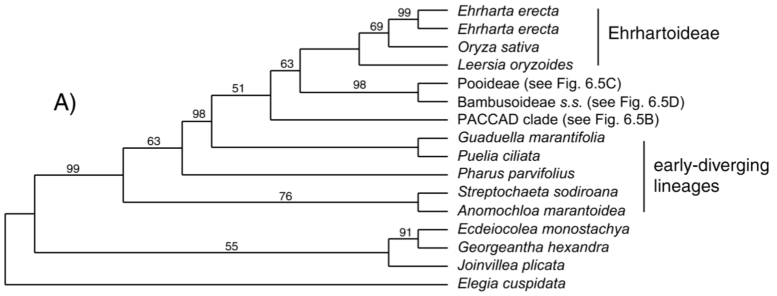


Figure 6.5A

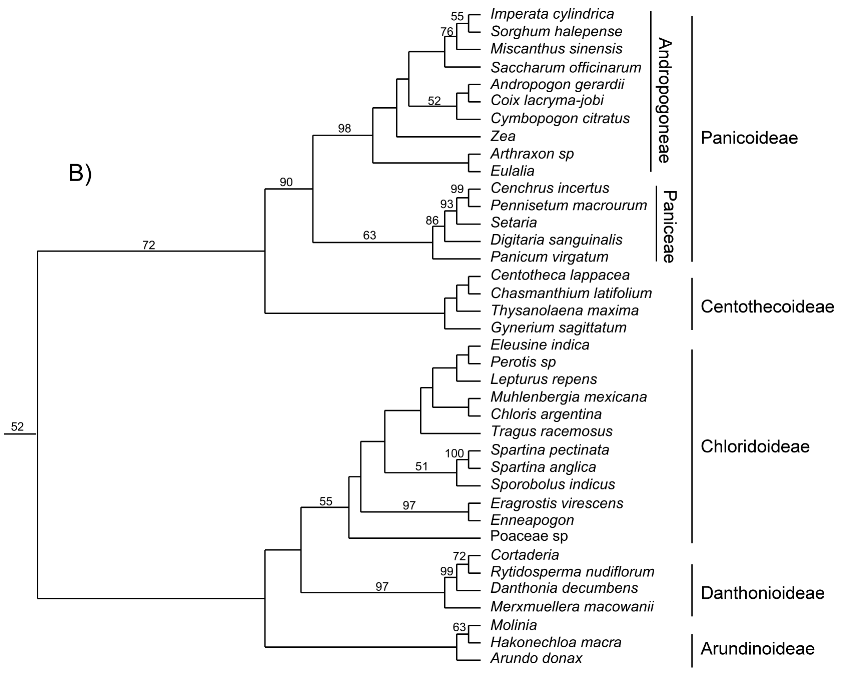


Figure 6.5B

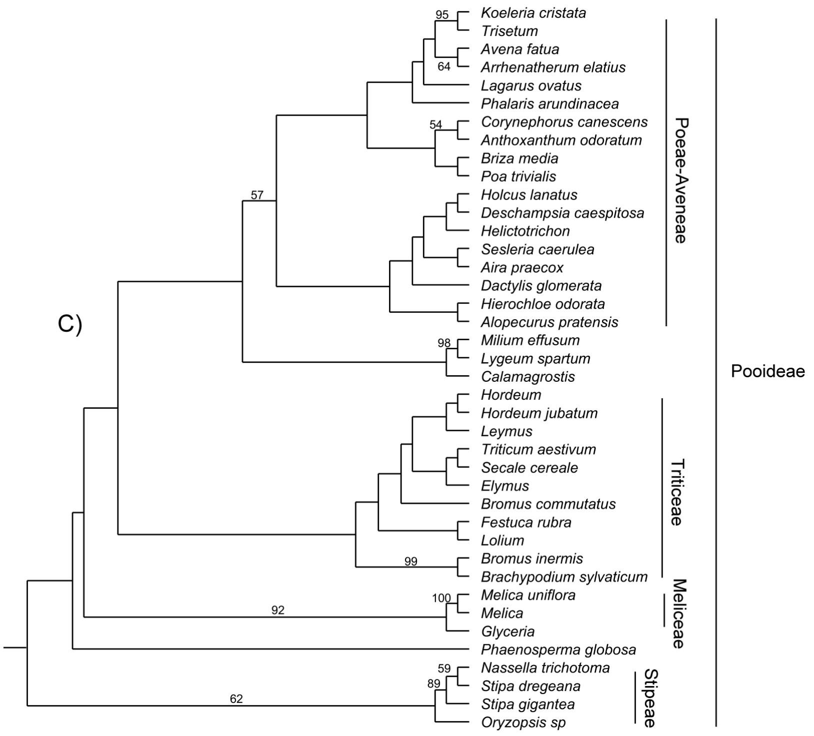


Figure 6.5C

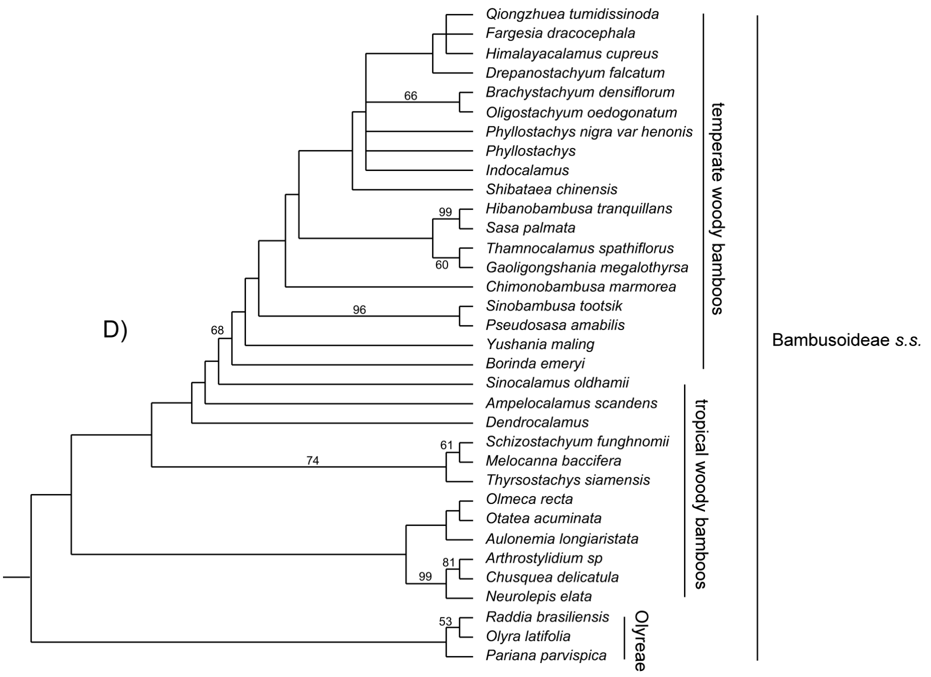


Figure 6.5D

Figure 6.5 - One of the equally most parsimonious trees obtained from MP analysis based on the combined *data set. Bootstrap support above 50% is shown on each branch. The tree is separated into four subtrees describing the major clades. A) basal relationships within Poaceae; B) PACCAD clade; C) Pooideae; D) Bambusoideae* s.s*.*

6.3.2 Divergence times

The divergence times obtained for the major clades of the grasses are shown in Table 6.1 and the tree with estimated ages based on the Bayesian analysis of the combined data set is presented in Figure 6.6. The results obtained with the different data sets were not always consistent, with some clades having similar divergence dates estimated from all data sets (e.g. Panicoideae, Andropogoneae or Chloridoideae; Table 6.1), while other clades obtained different estimates depending on the data set used (e.g. Triticeae, PACCAD clade or Ehrhartoideae; Table 6.1). The origin of the core Poaceae (i.e. excluding the early-diverging lineages) was dated between 49.8±6.2 and 41.6±4.6 Mya (Table 6.1), a relatively late appearance after the appearance of the last early-diverging lineages of grasses (~60 Mya; Fig. 6.6). Ehrhartoideae were dated between 36.1±5.4 and 23.0±12.2 Mya, and this wide range reflected the discrepancies concerning the possible relationships of the subfamily amongst Poaceae (Figs. 6.1 to 6.5). The origin of Pooideae was estimated between 38.8±5.8 to 30.8±1.5 Mya with all data sets (Table 6.1). The divergence times obtained for the PACCAD clade were similar to the one for the Pooideae, except with *trnLF* that gave younger estimates (23.6±13.5 and 24.5±9.6 vs. 31.7±2.6 to 33.4±7.9; Table 6.1). Within the PACCAD clade, Panicoideae and Chloridoideae were dated by all data sets to have originated between 19.1±9.3 and 25.0±7.2 Mya, while the other subfamilies of the clade being estimated as younger lineages (Table 6.1). The age given to Paniceae was close to the estimation obtained for the whole subfamily age, especially with the *trnLF* data set, while Andropogoneae were a younger lineage dated between 10.1±9,2 and 6.1±5.8 Mya (Table 6.1). Finally, the temperate bamboos were estimated as a much younger clade than the two other lineages within Bambusoideae *s.s.*, with ages of 2.0±1.9 to 12.1±4.6 Mya and 14.8±9.8 to 26.1±9.3 Mya respectively (Table 6.1).

The two time series corresponding to the past CO_2 levels between 65 and 52 Mya, and between 25 and 5 Mya were analysed to detect any changes in the CO_2 concentration. No significant change in slope could be detected at the 5% confidence level between 65 and 52 Mya (F=0.690; *p*=0.507), where the level of CO_2 remained high throughout the period. However, the period between 25 and 5 Mya showed a significant structural change in the time series (F=26.246; *p*<0.001) that was associated with the lowest concentration of CO_2 around 15 Mya (Fig. 6.7). This significant result indicated a new increase in CO_2 concentration from this point onwards, while the period before this point showed a decrease in those concentrations.

The two types of photosynthetic pathways were mapped on the phylogenetic trees obtained with all three data sets but only one is shown for clarity (Fig. 6.7). On all trees, the C_4 was hypothesised to have appeared twice during the evolution of the grasses, once on the branch leading to the Chloridoideae, and once on the branch leading to the Panicoideae, excluding *Gynerium*. The time associated with the first node possessing the C_4 metabolism were estimated as being between 19.6±9.5 and 25.0±7.2 Mya for the chloridoid ancestor, and between 17.2±6.4 and 20.8±9.3 Mya for the panicoid ancestor (Table 6.2). When these estimates were compared with the CO_2 data, the appearance of C_4 did not match the lowest value available, but was associated with a period of relatively low CO_2 level compared to other recorded time periods (Fig. 6.7).

Table 6.1 - Divergence times and confidence intervals in Mya for the main clades of the Poaceae. The names given to the different clades refer to Figs. 6.1 to 6.5.

Clades	*trnLF*		*rbcL*		combined	
	ML	Bayesian	ML	Bayesian	ML	Bayesian
PACCAD	24.5	23.6	33.3	33.4	32.3	31.7
	±9.6	±13.5	±7.1	±7.9	±4.6	±2.6
Panicoideae	19.9	19.1	19.9	20.8	22.3	21.7
	±7.8	±9.2	±8.8	±9.3	±3.9	±5.3
Andropogoneae	6.2	6.1	10.1	8.7	9.6	9.7
	±6.1	±5.8	±9.2	±8.6	±3.1	±2.7
Paniceae	19.1	18.4	14.5	14.9	15.6	15.0
	±9.3	±8.3	±11.3	±9.5	±2.6	±1.9
Arundinoideae	24.5	23.6	25.1	25.1	14.9	14.8
	±9.6§	±13.5§	±10.4	±11.4	±6.7	±4.9
Centothecoideae	17.8	17.1	20.7	21.6	19.9	20.6
	±9.9	±11.7	±13.1§	±8.3§	±5.7	±4.3
Chloridoideae	20.3	19.6	24.9	25.0	21.8	21.7
	±8.1	±9.5	±7.9	±7.2	±6.4	±3.1
Danthonioideae	14.7	14.3	22.4	22.4	22.3	22.0
	±14.3	±13.7	±8.1	±9.1	±4.3	±3.6
Pooideae	38.4	38.8	34.8	35.1	30.8	37.6
	±5.8	±5.8	±5.5	±7.2	±1.5	±1.4
Triticeae	12.4	12.1	11.7	12.1	29.3	24.4
	±8.1	±7.4	±10.1	±9.7	±3.9	±3.1
Poeae/Aveneae	25.2	24.9	15.1	15.4	37.4	28.4
	±6.5	±8.5	±11.6	±12.9	±4.6	±3.5
Meliceae	26.4	26.6	19.1	18.9	25.2	24.7
	±9.5	±10.1	±8.7	±11.5	±7.6	±4.8
Ehrhartoideae	30.9	31.7	23.0	23.1	26.0	36.1
	±11.4	±9.7	±12.2	±12.5	±4.9	±5.4
tropical woody bamboos	18.1	17.8	26.1	23.0	20.7	19.8
	±11.9	±9.9	±9.3§	±10.1§	±6.8§	±4.5
temperate woody bamboos	2.0	3.1	7.1±	6.3	12.1	7.9
	±1.9	±3.0	6.5	±4.6	±4.6	±2.9
herbaceous bamboos	22.9	22.8	14.8	16.2	24.0	22.1
	±10.5	±10.1	±9.8	±10.7	±1.8	±2.7
Poaceae (except early-diverging lineages)	42.9	41.6	44.5	44.8	49.8	47.3
	±5.9	±4.6	±5.6	±6.3	±6.2	±4.5

§clade was not monophyletic and the date of the oldest lineage of this clade was taken

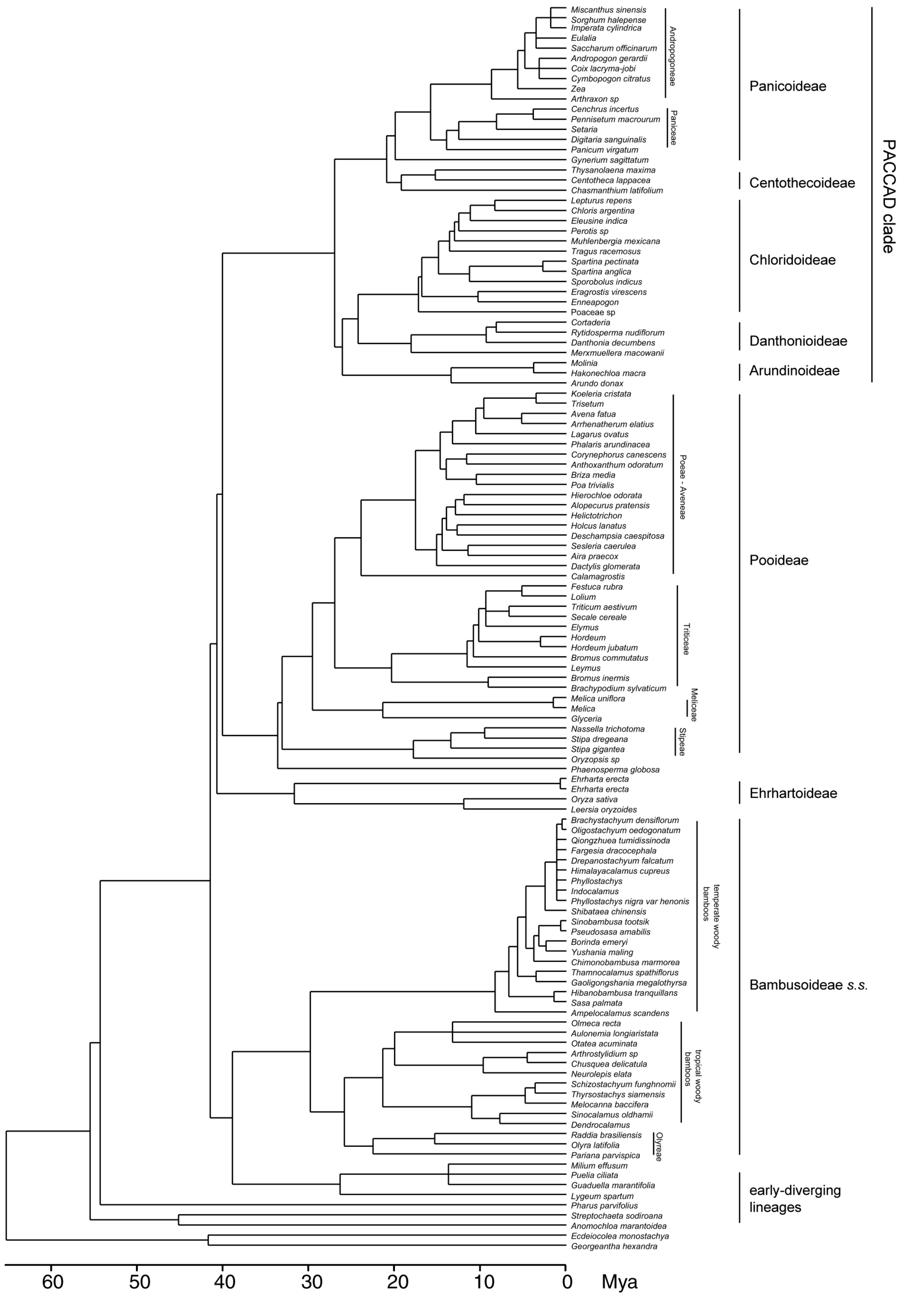


Figure 6.6 - Phylogenetic tree obtained with Bayesian analysis on the combined data set, and calibrated with the four fossils using the method of PL.

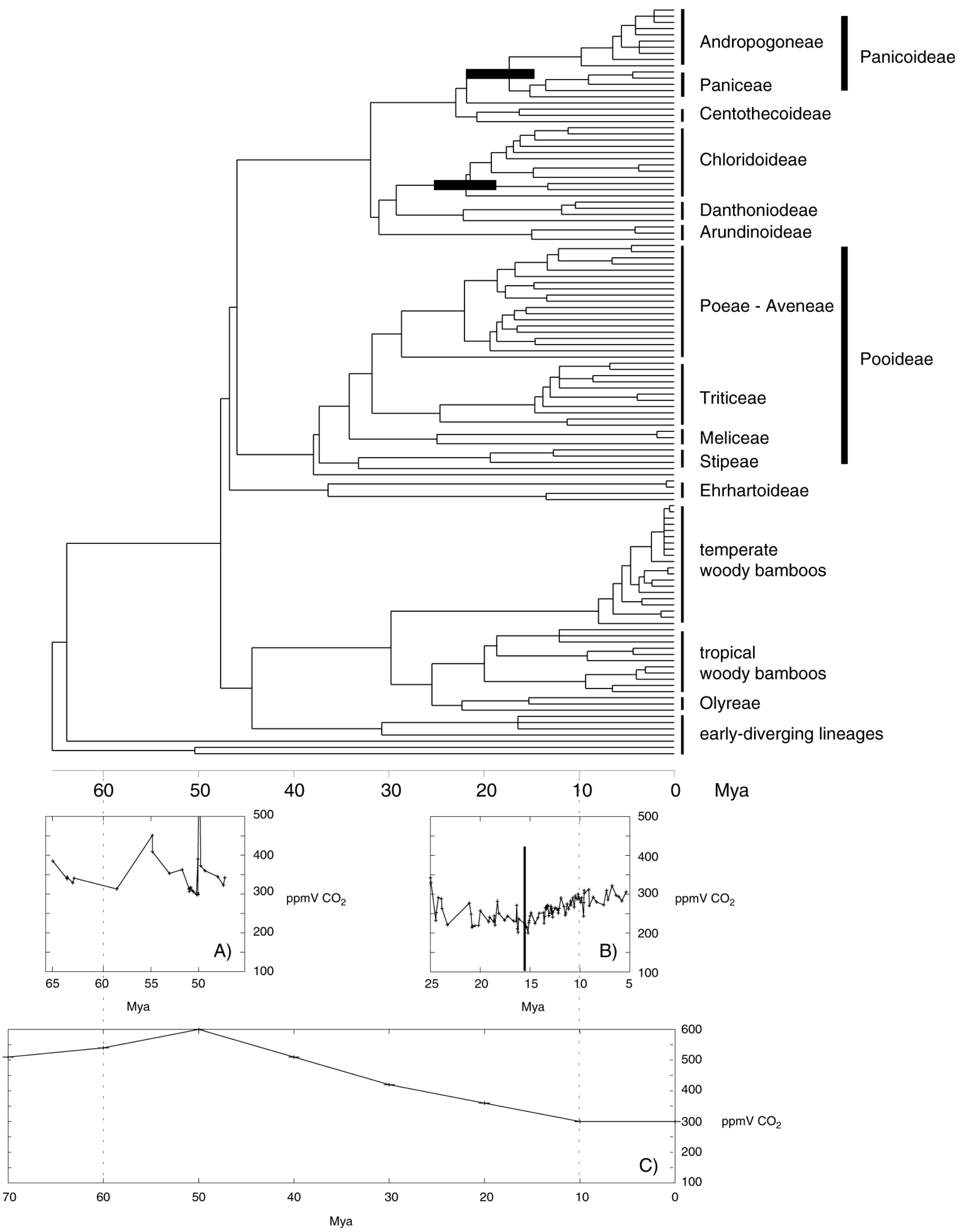


Figure 6.7 - Phylogenetic tree for Poaceae based on the Bayesian analysis of the combined *data sets, with the origin of the C_4 photosynthetic pathway. The horizontal black bar on the tree represents the estimated period of time corresponding to the appearance of C_4 grasses by taking the oldest and youngest estimate over all analyses. The estimated CO_2 levels are taken from A) Royer et al. (2001), B) Pagani et al. (1999) and C) Berner and Kothavala (2001). The vertical bar in B) indicates a significant change in the level of CO_2*

Table 6.2 - Origin of the C_4 photosynthetic pathway in the grasses. The dates correspond to the estimated age of the first node whose ancestral state was hypothesised as C_4.

Data sets	Analysis	chloridoid ancestor	panicoid ancestor
trnLF	ML	20.3±8.1	19.9±7.8
	Bayes	19.6±9.5	19.1±9.2
rbcL	ML	24.9±7.9	19.9±8.8
	Bayes	25.0±7.2	20.8±9.3
combined	ML	21.3±3.9	17.2±6.4
	Bayes	21.7±3.1	17.2±5.3

6.4 Discussion

6.4.1 Phylogenetic analyses

The phylogenetic trees presented in this chapter represent some of the largest evolutionary hypotheses proposed for the grasses. The structure of the trees obtained from the three different data sets is similar to previous phylogenetic studies concerning the whole family (e.g. Clark et al., 1995; Liang and Hilu, 1995; Duvall and Morton, 1996; Soreng and Davis, 1998; Hsiao et al., 1999; GPWG, 2001). Some of the controversies regarding relationships between major clades found between other data sets have also been highlighted in these phylogenetic trees. The *trnLF* and *rbcL* plastid regions gave different hypotheses for groupings within the family, but much of the well supported structure was consistent. For this reason, a combined analysis was performed. However, the bootstrap support for the clades in the combined analysis were very much reduced for groups within Poaceae, and the posterior probability were lower in some parts of the tree (Figs. 6.4 and 6.5). This could, for example, suggest incongruence between *rbcL* and *trnLF* regarding the monophyly of Bambusoideae *s.s.* or the placement of Pooideae.

The existence of the BEP clade is still an unresolved question, and the large sample of taxa used in the different analyses performed in this chapter could not give clear evidence to support it. It is also noteworthy that all Bayesian analyses placed Pooideae as sister to either the PACCAD or the PACCAD and Ehrhartoideae with high level of posterior probability (Figs. 6.1A, 6.3A and 6.4A). Model-based methods are less prone to statistical errors and can give more robust results than MP (Huelsenbeck, 1995; Swofford et al., 1996; Swofford et al., 2001), and thus could give more accurate estimates of the true grass evolutionary hypothesis under certain conditions. For MP, analyses based on *rbcL* also placed Pooideae as sister to the PACCAD clade (Fig. 6.3A), which is consistent with the results of the simulations found in chapter 5 and previous analyses of this plastid DNA region (Duvall and Morton, 1996; GPWG, 2001). The combined and the *trnLF* analyses placed Pooideae within a clade containing Bambusoideae and Ehrhartoideae (the BEP clade; Figs. 6.2A and 6.5A), which is a configuration found with previous studies on different DNA regions (e.g. Clark et al., 1995; Mathews and Sharrock, 1996; Clark et al., 2000; Mathews et al., 2000; Zhang, 2000). The branch lengths separating the different taxonomic groups forming the BEP and/or PACCAD-P clades remained very small (Figs. 6.1 to 6.5) and the estimated dates of divergence obtained suggested that the appearance of these grass lineages happened very quickly around 50 to 40 Mya (Table 6.1). Such a radiation could then explain the difficulties recent molecular studies have faced with the phylogenetic inference of the main grass lineages, and the question may remain difficult to resolve with certainty without the addition of much more data.

The Bambusoideae *s.l.*, as traditionally circumscribed, was a heterogeneous group including what is considered now as early-diverging grasses (e.g. *Anomochloa*, *Pharus*, *Puelia* and *Streptochaeta*; Clark et al., 1995), but recent molecular analyses have restricted the subfamily to a narrower group of woody and herbaceous bamboos (Clark et al., 1995; Duvall and Morton, 1996; GPWG, 2001). The herbaceous bamboos belong to Olyreae and are clearly separated from the other herbaceous bamboos making up the early-diverging lineages. In the present analyses, the monophyly of Bambusoideae *s.s.* is supported only with the *rbcL* data set analysed by Bayesian analysis, while the MP analysis showed low bootstrap support (Fig. 6.3D). The *trnLF* data set separated Bambusoideae *s.s.* in two paraphyletic groups, one containing the tropical woody and herbaceous bamboos (Fig. 6.1D and 6.2D) and one containing the temperate woody bamboos (Fig. 6.1E and 6.2E), although the Bayesian analysis was the only one to support these distinctions. The combined analysis was similar to *rbcL* alone, except that the posterior probability was much lower and no bootstrap support was associated with a monophyletic Bambusoideae (Figs. 6.4D and 6.5D). This could reflect incongruence between *rbcL* and *trnLF*.

Ehrhartoideae is a lineage strongly supported by molecular data (e.g. Clark et al., 1995; Duvall and Morton, 1996; Mathews and Sharrock, 1996; Soreng and Davies, 1998; GPWG, 2001) and is characterised by several morphological characters (e.g. one female-fertile floret per spikelet, inner whorl of stamens, styles not fused; GPWG, 2001). It was monophyletic in all analyses presented in this chapter, but its position within the grass family was ambiguous. The *rbcL* data set placed it as sister to the PACCAD clade and Pooideae, while *trnLF* positioned it within the BEP clade with MP or placed it as sister to the PACCAD clade with Bayesian analysis (Figs. 6.1A, 6.2A and 6.3A). Ehrhartoideae have been assigned to the BEP clade by previous studies (GPWG, 2001), and it is likely that the difficulties surrounding the resolution of the BEP clade are also affecting the placement of this lineage.

The monophyly of Pooideae was supported with the *rbcL* data set (Fig. 6.3C), and the Bayesian analysis performed on the *trnLF* and the combined matrices (Fig. 6.1C and 6.4C). However, no bootstrap support was associated with this clade in the MP analyses based on the two latter data sets, and low bootstrap values supported the core components of the subfamily composed of Triticeae, Aveneae, and Poeae (Figs. 6.2C and 6.5C). This is in contradiction to previous molecular studies including plastid DNA restriction site data (Soreng et al., 1990; Davis and Soreng, 1993; Nadot et al., 1994; Soreng and Davis, 1998). The low support found with *trnLF* could be due to the exclusion of characters associated with long gaps within the intron and intergenic spacer of this DNA region. With the Bayesian analyses, which showed good support within Pooideae, the position of the early-diverging Pooideae (e.g. Meliceae, Stipeae, *Phaenosperma*) was ambiguous between the analyses (Figs. 6.1C, 6.3C and 6.4C). This is also the case with other molecular studies where the order of divergence of these groups in comparison to the core Pooideae was not resolved (Soreng and Davis, 1998; GPWG, 2001). *Phaenosperma* was previously assigned to Bambusoideae based on morphological characters (Clayton and Renvoize, 1986), but molecular studies place it within Pooideae (GPWG, 2001). A particularly odd result obtained with the Bayesian analysis of the combined matrix was the placement of *Lygeum* and *Milium*, which were grouped with *Guaduella* and *Puelia* as sister to the Bambusoideae (Fig. 6.4D). This does not certainly reflect a biological relationship and the *trnLF* analyses, and previous molecular studies, have considered them as part of Pooideae within the small tribes Lygeae and Stipeae respectively (Figs. 6.1 to 6.2; Soreng and Davis, 1998; GPWG, 2001). However, the reasons of this grouping are unknown, but the introduction of missing characters for the sites corresponding to the *rbcL* sequence of *Lygeum* and *Milium* could have caused their odd placement in the combined tree.

The PACCAD clade is a well-defined groups in the Poaceae (Hamby and Zimmer, 1988; Doebley et al., 1990; Davis and Soreng, 1993; Nadot et al., 1994; Clark et al., 1995; Liang and Hilu, 1995; Duvall and Morton, 1996; Hsiao et al., 1999; Clark et al., 2000; Mathews et al., 2000; GPWG, 2001), and the support given by the present analyses to this clade was always high (Figs. 6.1 to 6.5). Besides molecular evidence, a series of morphological synapomorphies (such as elongated mesocotyl internode, loss of the epiblast or solid culm internodes; GPWG, 2001) seem to establish the monophyly of this clade. The relationships among the major lineages in the PACCAD clade are, however, not well established (GPWG, 2001). Panicoideae and Chloridoideae were supported by all three data sets, and the remaining

Aristidoideae, Arundinoideae and Danthonioideae were, when present, associated with Chloridoideae in a larger clade with the Bayesian analyses (Figs. 6.1B, 6.3B and 6.4B). These latter subfamilies were however not supported by the bootstrap analyses performed with MP (Figs. 6.2B, 6.3B and 6.5B). Centothecoideae were, however, associated with Panicoideae in all analyses with moderate to high support by the Bayesian and MP analyses, but low or no support was found to grant its status as a monophyletic subfamily (Figs. 6.1B to 6.5B). This confirmed the GPWG (2001) analyses that also found low bootstrap support for the monophyly of this subfamily. These findings are consistent with other recent molecular analyses and in particular with the GPWG (2001). Unfortunately, some taxa that were found difficult to assign to particular clades in previous studies are not present in the analyses presented in this chapter. For example, the genera *Micraira* and *Eriachne* have been a puzzle to grass systematists (GPWG, 2001), but either no sample could be obtained or their sequencing failed.

The position of the early-diverging grass lineages found in the analysis presented in this chapter reflects the ambiguities present in previous molecular studies (Clark et al., 1995; Hilu et al., 1999; Mathews et al., 2000; Zhang, 2000; GPWG, 2001). *Anomochloa* and *Streptochaeta* formed a monophyletic group with *rbcL* and the combined data sets, where both were present (Figs. 6.3A, 6.4A and 6.5A). This placement was similar to the large combined analysis of GPWG (2001) that concluded these should be combined into the subfamily Anomochloideae. However, when *Anomochloa* was not present in the analysis, *Pharus* was sister group to *Streptochaeta* (Figs. 6.1A and 6.2A). The sampling of these early-diverging lineages is sparse, and long branches are characteristic of the base of the grass phylogenetic tree. A denser sampling would probably help resolve these basal relationships, although the number of possible extant species to sample from is rather thin. For example, *Anomochloa* is a monotypic genus, while *Streptochaeta* has three species, and *Pharus* five (Clayton and Renvoize, 1986).

Finally, these analyses are inconclusive regarding the sister group of the Poaceae. The *rbcL* data set joined *Joinvillea*, *Ecdeiocolea* and *Georgeantha* into one sister group to the Poaceae, while with *trnLF Joinvillea* was sister to the Poaceae, and the combined analysis placed a group with *Ecdeiocolea*, *Georgeantha* (Ecdeiocoleaceae) as sister to the Poaceae (Figs. 6.1A to 6.5A). The monophyly of the grass family is however clear, either based on molecular or morphological characters (Doyle et al., 1992; Watson and Dallwitz, 1992; Clark et al., 1995; Soreng and Davis, 1998; APG, 1998; GPWG, 2001), and further work is needed to show conclusively which family from the Poales are sister to the Poaceae.

Many different DNA regions have been sequenced for phylogenetic studies of major clades (subfamilies) within Poaceae (e.g. *ndhF*, Clark et al., 1995; *rbcL*, Duvall and Morton, 1996; *matK*, Liang and Hilu, 1995; ITS, Hsiao et al., 1999; *rpoC2*, Barker et al., 1999; *rpl16*, Zhang, 2000; *phyB*, Mathews et al., 2000), but the analyses of the *trnLF* plastid region presented here have never been applied before at this level. The alignment of the *trnLF* region contained several long gaps that were diagnostic of certain clades. For example, the Andropogoneae are uniquely characterised by a 139 bp gap. However, the analyses performed did not take into account these long gaps as well as some smaller ambiguous ones as they were excluded from the phylogenetic analyses. The reason behind their exclusion was that maximum-likelihood based methods were used to analysed the data sets, or at least to estimate branch lengths, and current model of DNA evolution cannot incorporate gaps (Swofford et al., 1996). Nevertheless, the information contained in these gaps is certainly valuable and is being investigated further (Hodkinson et al., in prep.) using a gap-recoding strategy (e.g. Simmons et al., 2001).

6.4.2 Divergence times

The first unambiguous fossil evidence for the grass family have been found in the form of macrofossils dating from 65 to 55 Mya (Crepet and Feldman, 1991). However, the calibration of the phylogenetic trees obtained from molecular data sets suggested that the core Pooideae are more recent, with dates ranging between 49.8 and 41.6 Mya (Table 6.1), and that lineages leading to the major groups of grasses appeared within a few Mya (Figs. 6.7 and 6.8). This rapid diversification of ancestral grasses into the current main lineages could be the reason behind the difficulties of resolving the affinities between Bambusoideae, Ehrhartoideae, Pooideae and the PACCAD clade, and the numbers of different gene trees obtained as described in chapter 5 of this thesis (see also Nadot et al., 1994; Clark et al., 1995, 2000; Mathews and Sharrock, 1996; Hsiao et al., 1999; GPWG, 2001). The major diversification that led to the current 10,000 grass species seemed to happen even later as most of the major groups within the family only appeared around 30 Mya or later (Table 6.1; Figs. 6.7 and 6.8). This period of time corresponds to the appearance of most other ‘arid’ angiosperm families (Singh, 1988), and could suggest a major shift in grass evolution that allowed them to either acquire or make use of drought-tolerance characters such as increased root and decreased shoot growth, a general reduction in physiological activity in periods of drought, sunken stomata, and thick dense cuticles (Archibold, 1995; Willis and McElwain, 2002). These combined characteristics have been suggested to have conferred a competitive advantage to grasses in conditions of increasing global aridity that occurred during the Tertiary (Leopold and Denton, 1987; Wing and Boucher, 1998).

The standard deviation around the estimated times of divergence were generally large, with a time span of 15 to 20 Mya for the appearance of most groups. An interesting effect of the combined data sets was that the confidence interval were much more reduced than for *rbcL* or *trnLF*. However, the confidence intervals obtained by bootstrapping the data matrices does not reflect uncertainty in the topology, which will create another source of error. The effects of different topologies have been examined by Sanderson and Doyle (2001), and they are visible in these analyses for groups that were placed differently in the three analyses, such as Ehrhartoideae or the herbaceous bamboos for example.

6.4.3 CO_2 levels and C_4 photosynthesis

Over half the species of the grass family are included in the PACCAD clade (GPWG, 2001), and its two major subfamilies, Panicoideae and Chloridoideae, contain all C_4 grasses (Kellogg, 1999). The acquisition of the C_4 photosynthetic pathway could therefore have been a major key innovation in grass evolution. Further experimentations would be required to test whether the larger number of species possessing the C_4 pathway is due to that key innovation. A suitable approach would be at first to test if the number of species in those groups are larger than expected under a random model of speciation (see Barraclough and Nee, 2001 for a review). Two changes from C_3 to C_4 were hypothesised on all phylogenetic trees presented here (Fig. 6.7), but it is possible that the introduction of more taxa will increase the number of origins of C_4 as other panicoid tribes or genera from Paniceae, not sampled, are C_3 species (e.g. Isachneae and Neurachneae; *Canastra*, *Homolepis* and *Streptostachys*; Watson and Dallwitz, 1992; Chapman, 1996). The estimation of the age of the C_4 origin could therefore be overestimated in Panicoideae, with an older ancestor being assigned the first C_4 character state. This, however, is less likely to be the case for Chloridoideae where all but one species have been recognised as C_4 (Chapman, 1996; Kellogg, 1999). This C_3 exception is a species of *Eragrostis*, a genus of approximately 350 C_4 species within Chloridoideae, and is presumably a reversal. The C_4 pathway is present in several different alternatives within these two subfamilies, and a group of features would have to be in place in order for the C_4 photosynthesis to arise (Dengler and Nelson, 1999; Kanai and Edwards, 1999). Some genera such as *Panicum* are interesting because they establish evidence of intermediate forms and thus the possibility of a gradual transition from C_3 to C_4, as both C_3 and several form of C_4 are present in this genus (Chapman, 1996). By contrast, Chloridoideae have a relatively uniform C_4 photosynthesis, which can suggest a rapid evolution after the arrival of appropriate climatic conditions (Renvoize and Clayton, 1992; Chapman, 1996). The conditions that triggered the origin of C_4 in these two subfamilies are therefore unlikely to be identical. The carbon fixation in C_3 and C_4 photosynthesis is different, and evidence from the carbon isotopic composition of palaeosols (Cerling et al., 1997) and fossil tooth enamel (MacFadden and Cerling, 1994) have indicated that plants with C_4 photosynthesis evolved around 15 Mya, with a global expansion from 7 to 5 Mya (Cerling et al., 1993). The age estimates presented here indicated that the appearance of C_4 happened between 19.6 and 25 Mya for the chloridoid lineage and between 17.2 and 20.8 for the panicoid lineage (Table 6.2). The contradictions between the fossils dates and the molecular clock dates might result from the different estimates these dates provide. Estimates from tooth enamel and composition of palaeosols are likely to represent a period of time where C_4 grasses were abundant, while the dates obtained from the phylogenetic trees provide an estimate of the first appearance of the C_4 grasses. In that case, molecular-based estimates are expected to be older than fossil evidence. The molecular dates could also be a conservative estimates because they correspond to the nodes at which the diversification of the C_4 taxa occurred. The first C_4 plants would have occurred at some point along the branches leading to these nodes.

A decrease in past CO_2 levels has been hypothesised as a possible factor for the origin of the C_4 photosynthesis (Cerling et al., 1997; Pagani et al., 2001). The estimated origin of the C_4 from the molecular data does not correspond to the lowest concentration of CO_2 that is found around 15 Mya, but is still consistent with levels that were just above these lowest values (Fig. 6.7). Moreover, there has been a clear decrease in these concentrations since the appearance of the first grasses during the middle Eocene (~50-40 Mya), which could have put the first C_4 grasses at a selective advantage. C_4 plants are favoured under conditions of high temperature and low precipitation as they show a much more efficient CO_2 uptake and a reduced water loss in comparison to C_3 plants (Long, 1999). C_4 photosynthesis, however, is less efficient than C_3 photosynthesis in cold temperatures and temperate conditions (Larcher, 1995). The present-day distribution of grasses reflects these distinctions as C_4 grasses tend to be found in hot and arid regions, whereas C_3 grasses have a predominantly warm temperate to arctic distribution (Watson and Dallwitz, 1992; Chapman, 1996). It is therefore likely that an increase in higher latitude aridity and temperature in the low latitudes during the Miocene (~40-20 Mya; Cerling et al., 1997) leading to more arid climatic conditions coupled with a decrease in CO_2 levels have favoured the appearance and global expansion of C_4 grasses.

6.5 Conclusion

The phylogenetic relationships inferred from the *trnLF*, *rbcL* and combined data sets presented in this chapter were in agreement with previous molecular studies and indicated well supported groupings within the grasses at the subfamilial and tribal level. However, the difficulties surrounding the relationships amongst the major subfamilies was also apparent in these analyses, which confirmed the results of the simulations performed in chapter 5. The estimates of divergence times obtained for the family further indicated the rapid and recent apparition of the major subfamilies. This could explain the difficulties experienced during these reconstructions (and others) to position the major clades within the grass family. More data is clearly needed in order to resolve these relationships and to obtain a precise evolutionary hypothesis for the grass family. The estimation of divergence dates for the origin of the C_4 photosynthesis suggested that the low levels of CO_2 during the middle Miocene (~20-15 Mya) are correlated with the appearance of C_4 grasses and could have therefore been a factor in their appearance and expansion.

CHAPTER 7: CONCLUSIONS

The advent of molecular phylogenetics has opened a wealth of new possibilities to evolutionary biologists and the prospect of having accurate evolutionary hypotheses for many groups of organisms, is a stimulus for further research and development of new ideas. Improvements can come from technological advances *per se*, as well as from increased sophistication in the use of current methods. At the same time, comprehensive phylogenetic trees will allow a better understanding of evolutionary relationships and will serve as a basis for examining macroevolutionary processes affecting evolution. For example, insights into speciation processes can be better undertaken when most species are represented within a tree. Advances of molecular techniques offer a real possibility to obtain such comprehensive samples for many groups of organisms, and the association between computational power and theoretical development will certainly open even more possibilities.

The success and feasibility of reconstructing large phylogenetic trees, and some of their applications to investigate grass evolutionary history were examined in this thesis. For most groups of organisms building trees containing most of their diversity will require sampling hundreds or thousands of taxa. New methods such as supertrees (Constantinescu and Sankoff, 1995; Semple and Steel, 2000) or new algorithms to search the tree space (Farris et al., 1996; Lewis, 1998; Larget and Simon, 1999; Mau et al., 1999; Nixon, 1999; Quicke et al., 2001) have been proposed, but the simulations presented in chapter 2 indicated that actual methods currently used are not performing badly. Trees containing 567 taxa could be inferred with more than 90% of correct nodes with 3,000 bp of data only, and a 13,000 taxa tree was reconstructed with 80% of the nodes correctly inferred with 10,000 bp, which is encouraging given the problem faced when searching such a large tree space. Furthermore, when dealing with real data sets, support values for large phylogenetic trees can be obtained with less computational effort by using simpler heuristic options, which were shown in chapter 3 to give similar support values to more extensive searches. This was especially the case at higher levels of support.

Support values were also found important when integrated within the MRP supertree reconstruction. MRP, examined in chapter 4, is an efficient method to build comprehensive trees, as topologies, with limited overlapping sets of taxa, can be assembled into a single supertree. Although the relationships found within the grass family using supertree reconstruction methods were close to a method based on combined data sets of molecular sequences, MRP relies on searching through the tree space, which is time-consuming, and other supertree methods such as MinCut (Semple and Steel, 2000; Page, submitted), a polynomial-time algorithm, could play a major role in the reconstruction of a 'Tree of Life'.

However, even at lower taxonomic levels, comprehensive phylogenies are desirable. Simulations performed in chapter 5 on the grass family suggested that data sets containing less than 1% of grass species where impaired by stochastic error, and led to bias in phylogenetic reconstruction in the placement of wheat in regards to maize or rice. Increasing the sequence lengths alleviated the problem with most DNA regions investigated, but increasing the number of taxa was the solution to resolve the bias found in others. Although the addition of taxa in the simulations was specifically designed to break long branches, it is probable that a general increase in taxa sampling will help our understanding of grass phylogenetics. Grasses play a major role in human society, and an accurate and complete evolutionary hypothesis for the family is desirable. For instance, with the sequencing of the complete plastid genome and large parts of the nuclear genomes of several grass species, of which large proportions are already complete, reliable phylogenetic trees could help with the mapping of genes and the understanding of their function in other related species.

The results obtained in chapter 6 with two plastid DNA regions on an extended sampling of species confirmed some of the results found in recent molecular studies that have investigated the relationship of the major clades of grasses such as those containing wheat, maize and rice. The placement of wheat in relation to maize and rice was still ambiguous with MP, while Bayesian analyses consistently placed wheat with maize. Estimates of divergence dates suggested that the major grass lineages appeared rapidly between 40 and 30 Mya. This rapid diversification could certainly have posed problems to phylogenetic reconstructions. In such cases, model-based methods such as Bayesian analysis could be more capable of finding accurate evolutionary relationships, especially with limited number of characters (Swofford et al., 1996). Grasses have several different key morphological characters that allow them to survive arid conditions or grazing. Such specifications have certainly played a role in the importance taken by the grasses in many ecosystems on Earth. In chapter 6, investigation of past CO_2 levels and divergence times in grasses corroborated the idea that low CO_2 concentrations between 25 to 15 Mya could have triggered the origin and expansion of C_4 grasses. This particular photosynthetic pathway is represented in over half the grass species, and has certainly played a role in grass diversification.

CHAPTER 8. TECHNICAL NOTES

Bioinformatics are playing such a major role in biology that it is difficult to think of a biological study that would not need software and computer power in order to analyse its data. Phylogenetic methods are no exception and their use has increased with the availability of powerful computer architectures and software (Felsenstein, 2002). With the up-coming of genomics and the huge amount of molecular data that will be available in the next decades, as well as the development of more sophisticated methods to analyse data, the rise of bioinformatics as an important field in biology is almost certain.

The aims of this thesis focused on theoretical aspects of phylogenetic reconstructions, and a large component involved the use of large computer power and development of specific software. It is difficult to integrate this aspect within the previous chapters, as these are tools created in order to speed up the analyses or the deciphering of the results. At the same time, they allowed an improvement in my programming skills, which now include Java, C, and Perl languages among others. I am certain these skills will be valuable in the future and are worth all the debugging time spent! I also used some of my time during the three years of the project developing the management system for the DNA bank database of the molecular laboratory in the Botany Department of Trinity College. This can not be classified as 'academic' work as no pure research was involved, but I still want to include it as part of this thesis for at least the many long evening hours spent on creating the MySQL tables, trying to connect them in a meaningful way, and writing a usable Perl script to tackle the many possible queries (correct or not) any users could want to send to the database! I will ensure my DNA samples and those of others are continued to be utilised in the future.

For these reasons, I have decided to add this eighth chapter to the thesis. It is designed to present the different tools I have developed, explain their utility and function and provide the source code and the executable available on an attached CD-ROM. At the same time, each are downloadable[11] freely or accessible[12] from the Botany Department web site.

[11] http://www.tcd.ie/Botany/NS/software.html

[12] http://dnabank.bot.tcd.ie

8.1 Software

8.1.1 Java language

The software written in Java (Sun Microsystems) by myself called SuperTree version 0.85 and TreeCorrect version 1.2 are far more user-friendly than those written in other languages. This is in part due to the ease of implementing user graphical interface with Java, and more time was spent on their development. Both Java programs have a graphic interface allowing user to select the different options using menus and boxes and have been compiled for Unix flavour and Windows (9x, Me, NT and 2000) operating systems. There is no MacOS version as the Java runtime (version 1.3) used to run the software has not been ported to ‘classical’ Mac operating system (9.x and below), and I have not had the opportunity to compile them on MacOS X where the correct Java runtime is available.

SuperTree version 0.85

The SuperTree software implements the different coding procedures proposed for the MRP supertree reconstruction method (Baum, 1992; Ragan, 1992; Purvis, 1996; Ronquist, 1996). It has been used in chapter 4 to create supertrees for the grass family in order to compare these coding procedures. The software is still in development, and contacts have been made with Wayne and David Maddison to integrate the software within the Mesquite[13] package.

The software is intended to facilitate the transcription of phylogenetic trees containing overlapping sets of taxa into binary matrices usable by phylogenetic software such as PAUP*4b or PHYLIP3.5 (Felsenstein, 1996). A description of the coding procedures is given in chapter 4. The program works by reading source trees one by one from a text file. It then translates the internal edges into a structure containing all taxa descending from the node in consideration and the possible bootstrap percentage associated with it. When all trees have been read, a matrix is constructed from the structure held in the memory of the computer where each node becomes a column and each taxon a row. Depending on the options selected, the matrix, as well as the weighting, is then written into an output file.

The input file consists of a list of trees that have to be consistent with the NEWICK format. Either names or numbers can represent the taxa and support for each node can be added on trees where the information is available in the form of

[13] http://mesquiteproject.org/mesquite/mesquite.html

branch lengths. When numbers represent taxa, a conversion table can be added after the list of trees and the program will output matrices containing the corresponding names instead of the numbers. The output files are NEXUS or PHYLIP data files containing binary characters for all taxa present within the source trees, as well as a descriptive file containing information on each character. Note that due to the way PHYLIP handles character weighting (i.e. letters and digits are used to represent character weightings), no weighting is effectively added to the PHYLIP files.

TreeCorrect version 1.2

The TreeCorrect software can be used to compare a given model tree with trees saved in NEXUS format. The program has been used in chapter 2 to compute the percentages of angiosperm trees correctly recovered by MP and NJ, as well as in chapter 6 to compute the percentages of correct grass trees found by ML, MP, and NJ.

The software performs two different topological comparisons on the input trees and computes the different percentages by averaging over the number of trees present in one replicate (i.e. ‘begin trees’ section in the NEXUS format) and over all replicates found in the input file. The trees and model tree can contain polytomies, and the model tree can consist of only a subset of the taxa found in the NEXUS trees. In such case, only the splits contained in the model tree will of course be considered. Firstly, it compares whether each internal edge from the model tree defined the same splits in each tree saved in NEXUS format. A tree is therefore either identical to the model tree or not, which leads to the percentage of correct trees. Secondly, it computes how many splits are identical between the model tree and each tree saved in NEXUS format. Within each replicate, two possibilities appear when multiple trees are present leading to two different percentages of trees correct. A split can either be considered correct if and only if it is present in all trees from a replicate or the percentage of tree correct is calculated on each tree of a replicate and the average is taken.

Two input files are required for the program to run. Firstly, a model tree file has to be open that contains one or several trees (or subtrees). Subsequently, the model tree(s) can be created directly within TreeCorrect using a function to interactively build NEWICK trees. Secondly, a NEXUS tree file containing the different replicates has to be inputted into the program. The output file saved by TreeCorrect consists of the three different percentages described above, as well as a detailed description of the percentage of time each edge from the model tree is present in the NEXUS tree file. A ‘verbose’ option is also available allowing a longer description of each edge containing details from each replicate rather than a summary over all replicates.

8.1.2 C language

Three other programs have been written in C that can perform various manipulations of phylogenetic trees. They are command driven software with no user-friendly graphic interface (being their only user so far I did not spend time developing the user interface). Therefore, input files or commands that do not correspond exactly to what the software expect will result in a crash without other explanations. However, ReadMe files can be found in the tar or zip balls on the web site or on the CD-ROM that should give enough information to use these programs without too much effort. I plan to improve these programs and to fuse them into one user-friendly package and to include some of the Perl script found on the CD-ROM.

GenTree version 0.5

GenTree is used to generate topologies according to a Yule or ‘pure birth’ process (see Steel and McKenzie, 2001 for a mathematical description). Branch lengths are assigned according to a Gamma distribution with shape given by the user and with scale of one. The different parameters required by the program are entered interactively by the user after the launch of GenTree. They consist of a number of taxa, the shape of the Gamma distribution, a restriction parameter for the branch lengths and standard deviation around the parameter and an output file name. The restriction parameter entered by the user allows the daughter branch lengths to be constrained within a certain range from the parent branch length. The maximum and minimum lengths allowed are calculated as the ratio of the daughter branch length over the parent one plus or minus the standard deviation given by the user. For example, a value of zero will put no restriction on the daughter branch lengths while a value of one will force the daughter branch lengths to be equal to the parent one.

The steps used by the program are as follows:

1. Starting with two edges with branch lengths drawn from a Gamma distribution, randomly decide which edge will be split into two new taxa. At each cycle of the loop, all edges have the same probability to give birth to a new taxon.

2. Assign a branch length for the two new daughter branches from the Gamma distribution.

3. Check whether the ratio parent/daughter branch length is within the value specified by the user. If not, go back to 2.

4. Go back to 1. until the number of taxa required is reached

An additional option that will be added soon to the program is to allow the user to specify a certain value of imbalance (Fusco and Cronk, 1995) that the generated tree should have. The program has been used in chapter 2 to create the 13,000 taxa tree used for the simulations.

BranchCut version 0.3

BranchCut is a very simple program that breaks long branches in a tree by adding new taxa. The different parameters required by the program are entered interactively by the user after the launch of BranchCut. The input parameters are an input and output file name, the number of taxa present in the original tree and the number of taxa that will be present in the output tree. The last parameter is a real number between zero and one that represents where the new branch will be inserted onto the existing branch. A value of zero means that a polytomy should be created at the base of the existing branch, while a value of one means that a polytomy should be created at the top of the existing branch. Intermediate values will cut the existing branch proportionally, for example, 0.2 will split the existing branch at 1/5th of its length and create a new node leading to two descendants each with a length of 4/5th of the previous existing branch.

The steps used by the program are as follows:

1. Select the longest branch in the tree (if more than one have the same length, take the first one found)

2. Split the branch selected in 1. into two descendants and assign the new branch lengths according to the parameter entered by the user.

3. Go back to 1. until the number of taxa specified is reached.

BranchCut simply takes a NEXUS tree as input and returns a larger tree as output. The format of the input file has to be rather rigid, as the tree must be found on the first line in the input file with no characters before the first opening parenthesis defining the NEXUS tree. Moreover, the taxa must be represented by numbers and not by species names. The program has been used in chapter 5 of this thesis and in Savolainen et al. (2002).

RandomTaxa version 0.1

RandomTaxa is a program to create replicates of a given PAUP*4b matrix by randomly sampling without replacement a fraction of the taxa. The different parameters required by the program are entered interactively by the user after the launch of RandomTaxa. When started, the program will ask the user to enter an input file name that has to be a NEXUS file containing a data matrix and an output file name. However, the headers defining the NEXUS sections within the input file have to be in lower case. At the same time, a tree has to be appended within a 'begin trees' section after the matrix with a translation table. Then, the program requires the number of taxa that will be randomly sampled from the original matrix. Accepted values are between one and the number of taxa found in the original matrix. Finally, the number of replicates desired have to be entered and the name of a file containing the commands that are used to execute each replicate. This 'begin paup' section will be appended after each replicate created and will be executed by PAUP*4b after each matrix randomly created.

RandomTaxa has not been used directly in chapters included in this thesis, but was written to analyse a sample of the angiosperm matrix used in Savolainen et al. (2002).

8.1.6 Perl scripts

Several Perl scripts have been written in order to manipulate files or trees and to analyse the results obtained in the different chapters. All these scripts are easy to use and to understand directly by reading the Perl script itself. There is therefore no need to describe all of them here and they have been put onto the CD-ROM attached.

8.2 DNA bank

A DNA bank has been created by Trevor R. Hodkinson to store and keep DNA samples extracted by researchers in the Molecular Laboratory of the Botany Department of Trinity College for further studies. Besides the need for the physical storage of the samples, it was decided that an information retrieval system should be implemented to allow easy and broad access to the information associated with the DNA samples stored. Due to the limited resources available for the creation and maintenance of the database, an obvious solution was to take advantage of the numerous open source software existing. Besides being free, they represent a very stable and solid alternative to the often expensive commercial counterparts, especially for a medium size database such as the Botany DNA bank. The database has been installed on a Dell Optiplex GX1 computer (Intel Pentium III; 600 MHz; 256 Mb RAM; 20 Gb hard disk) running SuSE7.3 Linux operating system. The Linux operating system is a free twin of UNIX that has, among other things, similar capabilities in terms of networking and database management, and represents an ideal solution for the Botany Department.

The MySQL[14] engine version 3.22.32 was chosen as the Database Management System (DBMS) as it is freely available, efficient and sufficiently flexible for the needs of the laboratory database. MySQL is a relational database in which the collection of data is organised into tables, which are then further divided into columns representing the different attributes of the data, and rows that gather the data itself. The DNA bank database currently has 31 different tables (with 36 fields) corresponding to different items of information such as plant family, genus, author of extraction, and type of DNA marker generated from the DNA. Important passport data such as collection locality, rarity, and habitat type can also be accommodated. It will, when fully complete, be compatible with the International Transfer Format for Botanic Gardens, Record Plant (ITF) version 2.

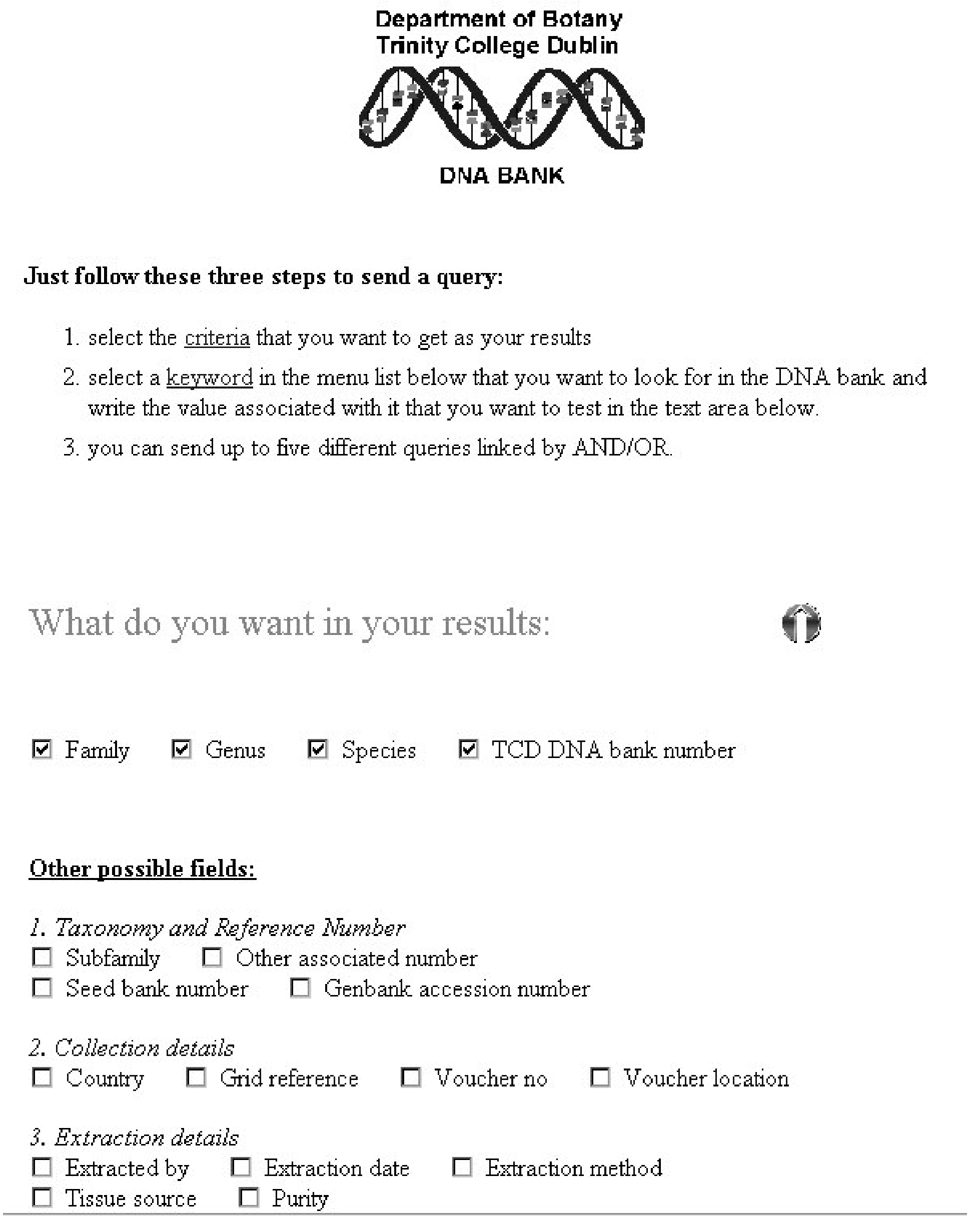
MySQL queries can be difficult to interpret and create for users that are not familiar with database systems. As the goal of the database was to allow the widest range of people to access the information stored in the DNA bank, a user-friendly HTML interface was created allowing the user to select the amount of information they wish to receive as results (Fig. 8.1) and to send queries to the database (Fig. 8.2). The design followed well-known databases such as SRS[15] that have become familiar to most molecular biologists. Therefore, instead of interacting directly with the DBMS the user operates the database through the World Wide Web, making it easily and widely accessible. The Apache web server version 1.3.12[16] was chosen as the software allowing us to transform the Dell computer into an Internet server and the address dnabank.bot.tcd.ie was chosen as the URL. More information concerning the database, its fields and operation, can be found on the help page associated with the DNA bank[17].

[14] http://www.mysql.org

[15] http://srs.ebi.ac.uk

[16] http://www.apache.org

[17] http://dnabank.bot.tcd.ie/help.html



Department of Botany
Trinity College Dublin

DNA BANK

Just follow these three steps to send a query:

1. select the criteria that you want to get as your results
2. select a keyword in the menu list below that you want to look for in the DNA bank and write the value associated with it that you want to test in the text area below.
3. you can send up to five different queries linked by AND/OR.

What do you want in your results:

☑ Family ☑ Genus ☑ Species ☑ TCD DNA bank number

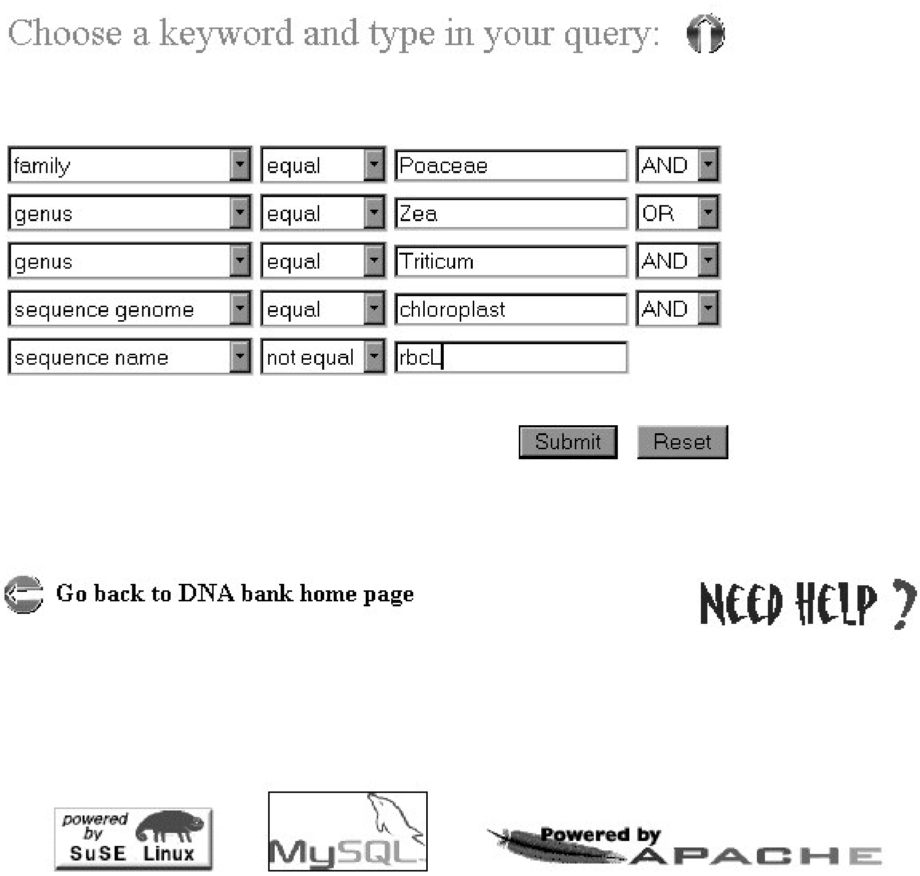
Other possible fields:

1. Taxonomy and Reference Number
□ Subfamily □ Other associated number
□ Seed bank number □ Genbank accession number

2. Collection details
□ Country □ Grid reference □ Voucher no □ Voucher location

3. Extraction details
□ Extracted by □ Extraction date □ Extraction method
□ Tissue source □ Purity

Figure 8.1 - Snapshot of the HTML page used to submit the MySQL queries. Users are asked to define the kind of information they want to appear in the result page. Basic search only include the family, genus, species and the TCD DNA bank number of the queried DNA sample, but user can customise the search to include many fields.



Choose a keyword and type in your query:

family	equal	Poaceae	AND
genus	equal	Zea	OR
genus	equal	Triticum	AND
sequence genome	equal	chloroplast	AND
sequence name	not equal	rbcL	

Figure 8.2 - Snapshot of the HTML page used to submit the MySQL queries. Users are asked to define the query they want to submit to the DBMS. Five different queries can be combined using AND, OR or NOT to tailor the needs of all users. Each query is represented by a keyword defining a table in the database, and a field entered by the user that corresponds to the information required.

The link between the web page seen by the user and the MySQL tables consists of a Perl script. The function of the script is to translate the queries entered on the HTML page into MySQL statements, send these statements to the MySQL server running on the Dell computer which will then return the results to the same Perl script. The script will finally format this information into an HTML page readable by the user. For more details, please refer to the Perl script itself found on the CD-ROM attached.

GLOSSARY

Bayesian analysis – Bayesian analysis combines a prior belief about the probability of a hypothesis with the likelihood of that hypothesis. The likelihood represents the information about the hypothesis contained in the observed data, and is identical to the likelihood function used in ML[18]. The method is based on the Bayes theorem, which when applied to phylogenetics will calculate the posterior probability of a phylogenetic tree τ . The posterior probability of tree τ_i , conditioned on the observed matrix of aligned DNA sequences, X , is obtained using Bayes formula

$$f(\tau_i \mid X) = \frac{f(X \mid \tau_i) f(\tau_i)}{\sum_{j=1}^{T} f(X \mid \tau_j) f(\tau_j)}$$

The likelihood of the i th tree is $f(X \mid \tau_i)$ and the prior probability of the i th tree is $f(\tau_i)$. The summation in the denominator is over all possible trees for s species (Yang and Rannala, 1997; Mau et al., 1999; Larget and Simon, 1999). Typically, an uninformative prior is used for trees, such that each has the same probability to be the correct tree. If prior evidence favoured some of the trees, the prior could be changed to give higher prior probability to such trees.

Distance methods – The use of corrected distances to account for superimposed changes at a single sequence position is an alternative to the use of likelihood for minimizing the impact of underestimation of the true amount of changes in a DNA sequence. The corrected distances are an estimate of the true evolutionary distances, which reflects the actual mean number of changes per site that have occurred between a pair of sequences since their divergence. Two categories of methods can then be applied on the transformed DNA sequence. One assumes the property of tree additivity (i.e. the evolutionary distance between each pair of taxa would be equal to the sum of the lengths of each branch lying on the path between the members of each pair), and that distances satisfy the four-point metric condition (Buneman, 1971):

$$d_{AB} + d_{CD} \leq \max(d_{AC} + d_{BD}, d_{AD} + d_{BC})$$

[18] Please, refer to *Maximum Likelihood* for more details on the calculations required to compute the likelihood of a tree.

where d_{ij} is the distance between taxa i and j, and $\max()$ is the maximum value function. Tree-additive distance can be fitted to an unrooted tree such that all pairwise distances are equal to the sum of the lengths of the branches along the path connecting the corresponding taxa. Due to the finite amount of data available, random error will cause deviation of the estimated evolutionary distances from perfect tree additivity. The idea is therefore to attempt to optimise an objective function that quantifies the degree of distortion between the path length and observed distances. The Fitch-Margoliash and related methods use the function

$$E = \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} w_{ij} \left| d_{ij} - p_{ij} \right|^{\alpha}$$

where E defines the error of fitting the distance estimates to the tree, T is the number of taxa, w_{ij} is the weight applied to the separation of the two taxa, d_{ij} is the pairwise distance estimate, p_{ij} is the path length connecting the two taxa on the given tree and α equals 1 or 2 (Swofford et al., 1996). The minimum evolution methods also use the previous equation to fit the branch lengths, but evaluate and compare trees with a different criterion:

$$LS\ length = \sum_{k=1}^{2T-3} \left| v_k \right|$$

where v_k is the branch lengths that minimise the sum of squared deviations between observed and path-length distances (Swofford et al., 1996).

The second approach is to use methods assuming ultrametric distances. Ultrametric distances are defined by satisfaction of the three-point condition

$$d_{AC} \leq \max(d_{AB}, d_{BC})$$

which will fit a tree so that the distance between any two taxa is equal to the sum of branches joining them, and the tree can be rooted so that all of the taxa are equidistant from the root (Swofford et al., 1996).

Error, bias and inconsistency - From a statistical point of view, systematic error is define as deviation between a parameter of a population and an estimate of that parameter, due to incorrect assumptions in the estimation method. Systematic error persists and may intensify as sample size increase and become infinite. Random error is defined as deviation between a parameter of a population and an estimate of that parameter, due strictly to a limited sample size used to make the estimate. By definition, random error disappears in infinite samples.

In a phylogenetic context, systematic error will occurs when the evolutionary process violates the assumptions of a phylogenetic method in a critical way. Under such conditions, a bias may be introduced favouring some branching patterns and decreasing the support for others, which could overcome the legitimate support for the true tree. As shown for parsimony and compatibility methods by Felsenstein (1978), because the effect is systematic, the addition of more data will increase the probability of recovering the incorrect tree. The method is then said to be inconsistent or positively misleading (Felsenstein, 1978a). All methods are consistent when their assumptions are met, but some methods such as ML are more robust to violations of their assumptions.

However, even if the evolutionary process underlying the data follows the assumptions of a particular analytical method, an incorrect tree may be inferred with finite data due to chance alone, which introduce random error. Random error does not necessarily produce a random effect on the outcome of the analysis. For example, for many methods of calculating pairwise distances, small distances and large distances are affected differently by sampling error (Swofford et al., 1996; Hillis et al., 1994).

Exact search strategy – Exhaustive search algorithm works by starting with three taxa, and then adding the ith taxon to each branch of every tree containing $i-1$ taxa generated during the previous step, keeping all trees in memory. The number of unrooted trees for T taxa is given by

$$B(T)=\prod_{i=3}^{T}(2i-5) \text{ (Felsenstein, 1978b)}$$

and it becomes unfeasible to enumerate all possible trees for more than 10 or 11 taxa.

Branch and bound algorithm is an exact search strategy that does not require exhaustive enumeration of all possible trees. In this procedure, the tree-space is search in a similar way as for the exhaustive search, but an upper limit for the tree length is set (e.g. by taking a random tree) and addition of taxa that makes the length of the growing tree exceeds this limit is prohibited and that tree is discarded (Swofford et al., 1996). The branch and bound algorithm is still extremely time-consuming, and data sets containing more than 20 or 30 taxa are too large for this algorithm to be used.

***Heuristic search* –** Heuristic search algorithms sacrifice the guarantee of optimality in favour of reduced computing time, and they actually represent the only way to infer phylogenetic trees with moderate to large number of taxa. Heuristic searches generally are hill-climbing methods. An initial tree is created and the algorithm will try to improve its score under the chosen optimality criterion by rearranging it using branch-swapping algorithms (e.g. NNI, SPR or TBR). The most common method to obtain the starting tree is by stepwise addition of taxa into a growing tree. However, the decision as to which taxa is to be added and at what time is far from being straightforward, and current software implement different strategies or left the option to the user. Unfortunately, there seems to be no strategy that works best for all data sets, and the empirical approach to try as many alternatives as possible is often the best (Swofford et al., 1996). An alternative to the stepwise addition is the star decomposition method that works by starting with a 'star tree' containing a single internal node. All possible trees constructing by joining two terminal nodes are then evaluated, and the best tree is saved for the next step until a binary tree is obtain (Swofford et al., 1996).

The major problem of such algorithms is that it is impossible to know whether the best tree found represent a global or local optimum, and several methods have been proposed to reduce the possibility of being trapped in such a local optimum. The idea is to allow the algorithm to go downhill and accept a suboptimal solution with a small probability in order reach other islands of trees (Swofford et al., 1996).

***Markov chain Monte Carlo* –** Due to the necessity to integrate over all possible trees in the denominator of the Bayes formula, the posterior probability of phylogenetic trees cannot be calculated analytically. The posterior probability can however be approximated by sampling trees from the posterior probability distribution using Markov chain Monte Carlo algorithms (Yang and Rannala, 1997; Mau et al., 1999; Larget and Simon, 1999), such as the Metropolis-Hastings-Green algorithm (MHG; Metropolis et al., 1953; Hastings, 1970; Green, 1995) or the Metropolis-coupled Markov chain Monte Carlo (MCMCMC; Geyer, 1991). The MHG algorithms work by constructing a Markov chain that has as its stationary frequency the joint posterior probability of the tree τ , the branch lengths v , the parameters of the model of substitution θ, and the rate heterogeneity α, if taken into account. These represent the state Ψ of the chain. The Markov chain is initialised by, for example, randomly picking a state from the prior. At each generation, a new state Ψ' is proposed and it is accepted with probability

$$R = \min(1, \frac{f(X \mid \Psi')}{f(X \mid \Psi)} \times \frac{f(\Psi')}{f(\Psi)} \times \frac{f(\Psi \mid \Psi')}{f(\Psi' \mid \Psi)})$$

A uniform random variable between 0 and 1 is drawn, and if this number is less than R the proposed state is accepted and $\Psi = \Psi'$ (Huelsenbeck and Ronquist, 2001). The MCMCMC variant runs several chains, some of which are heated by a certain factor, which means that the posterior probability of a tree is raised to some power β. Heated Markov chains can better explores the parameter space by allowing crossing of valleys within this landscape. The MCMCMC algorithm runs each chain for one generation, and swaps the states of the different chains according to a probability function (Huelsenbeck and Ronquist, 2001).

The Markov chain is usually run for several hundreds thousand generations, with state samples taken every so often. The samples from the Markov chain form a valid but dependent sample from the posterior probability distribution (Tierney, 1994).

Maximum likelihood – ML methods evaluate a phylogenetic tree in terms of the probability that a proposed model of substitution and the hypothesized history would give rise to the observed data. Put more formally, if given a data D and a hypothesis H, the likelihood of that data is given by

$$L(D) = P\{D \mid H\} \text{ (Edwards, 1972)}$$

It is important to remember that unlike probabilities, likelihood do not sum to one. Given a tree and a model of substitution, the probability of obtaining all possible data sets could be estimated, which will sum to one. However, only the probability of obtaining the observed data set is of interest here.

To calculate the likelihood for some site j, all possible scenarios by which the tip sequences could have evolved must be considered, and a probabilistic model for the process of nucleotide substitution is required to 'weigh' these different scenarios. For an ancestor A with two daughter sequences B and C, the likelihood for position j is

$$L_{(j)} = \sum_m \pi_m \times [\sum_k P_{m,k}(v_{AB}) L(x_{Bj} = k)] \times [\sum_k P_{ml}(v_{AC}) L(x_{Cj} = l)]$$

where v_{xy} is the length of the branch joining x to sequence y, π_x is the equilibrium frequency of nucleotide x and $P_{x,y}$ is the probability of substituting x by y. The second term on the right-hand side is the probability of state i changing to state k in the interval v_{AB}, $P_{ik}(v_{AB})$, times the likelihood that sequence B has state k at the corresponding position. If B is a known sequence, the likelihood will be one, whereas if B is an unknown sequence (i.e. internal node), then the likelihood of it having state k are derived recursively by inserting another copy of the right-hand part into the equation (Felsenstein, 1981; Swofford et al., 1996).

Under the assumption that the nucleotide patterns are independent, the likelihood for each pattern is combined into a total value for the whole sequence. Other parameters such as the rate of heterogeneity for each site can also be incorporated by modifying the probability of change from nucleotide i to j, $P_{i,j}$ (e.g. Yang, 1993; Steel et al., 1993).

Maximum parsimony – MP operates by selecting trees that minimize the total number of evolutionary steps required to explain a given set of data. The general MP problem can be defined as follows. From the set of all possible trees, find all trees T such that

$$L(T) = \sum_{k=1}^{B} \sum_{j=1}^{N} w_j * diff\,(x_{k'j}, x_{k''j})$$

is minimal, where $L(T)$ is the length of tree T, B is the number of branches, N is the number of characters, k' and k'' are the two nodes incident to each branch k, $x_{k'j}$ and $x_{k''j}$ represent either elements of the input matrix or optimal character-state assignments made to internal nodes, and $diff\,(y,z)$ is a function specifying the cost of a transformation from state y to state z along any branch. The coefficient w_j assigns a weight to each character. Note that $diff\,(y,z)$ needs not to be equal to $diff\,(z,y)$ (Swofford et al., 1996).

MP analysis actually comprises a group of related methods differing in their underlying evolutionary assumptions. The two simplest methods imposed no (Fitch, 1971b) or minimal (Wagner; Kluge and Farris, 1969; Farris, 1970) constraints on permissible character-state changes and both are appropriate under the assumption that probabilities of character change are symmetrical. Dollo parsimony is more appropriate when this symmetry is not expected (Farris, 1977; DeBry and Slade, 1985). The method of Camin and Sokal (1965) makes the strongest assumption of any of the methods, namely, that evolution is irreversible. All MP variants can be represented by a generalized method that assigns a cost for the transformation of each character state to the other possible states (Sankoff, 1975; Sankoff and Rousseau, 1975). The costs may be represented by a $m \times m$ matrix S, where S_{ij}

represents the increase in tree length associated with a transformation from state i to state j, and m is the total number of possible states.

NJ – NJ is the most commonly used star decomposition method, and is conceptually related to cluster analysis, but removes the assumption of ultrametricity of the data. However, tree additivity is assumed by NJ, and correcting for superimposed substitutions is important when using N.

The original distance matrix is modified by adjusting the distance between each pair of taxa in function of their average divergence from all other taxa. This modification has the effect of normalizing the divergence of each taxon for its average clock rate, thus removing the assumption of ultrametricity (Swofford et al., 1996). The tree is constructed by joining the least-distant pair of taxa i and j in the modified matrix, and their common ancestral node u is added to the tree. The two terminal taxa i and j are removed from the matrix, and the distance from the newly created node u to each other taxa k in the matrix is calculating by

$$d_{ku} = (d_{ik} + d_{jk} - d_{ij}) / 2$$

The process is repeated until two nodes remain, separated by a single branch, thus creating a binary tree.

NNI – branch-swapping by nearest-neighbour interchange. Each internal branch of the tree defines four subtrees, two on each side of the branch. NNI consists by interchanging one subtree on one side of the branch with one of the two from the other side. There are therefore only two rearrangements possible, and NNI is the quickest branch-swapping algorithm.

SPR – branch-swapping by subtree pruning and regrafting. SPR consists of pruning one subtree from the tree, which is subsequently regrafted to a different location on the tree. All possible subtree removals and reattachment points are evaluated. SPR is less time consuming than TBR but more than NNI.

TBR – Branch swapping by tree bisection and reconnection. TBR consists of dissecting the tree into two subtrees, which are then reconnected by joining a pair of branches, one from each subtree. All possible bisections and pairwise reconnection are evaluated. TBR is the more computationally intensive branch-swapping algorithm, but also the one that covers the widest range of trees from the tree-space.

BIBLIOGRAPHY

ANDERSON, J. S. 2001. The phylogenetic trunk: Maximal inclusion of taxa with missing data in an analysis of the lepospondyli (Vertebrata, Tetrapoda). Syst. Biol. 50:170-193.

ANDREWS, D. W. K. 1993. Tests for parameter instability and structural change with unknown change point. Econometrica 61:821-856.

ANDREWS, D. W. K. AND W. PLOBERGER. 1994. Optimal tests when a nuisance parameter is present only under the alternative. Econometrica 62:1383-1414.

APG. 1998. An ordinal classification for the families of flowering plants. Ann. Missouri Bot. Gard. 85:531-553.

ARCHIBOLD, O. I. V. 1995. Ecology of world vegetation. Chapman and Hall, London.

ARNQVIST, G. AND D. WOOSTER. 1995. Meta-analysis: synthesizing research findings in ecology and evolution. Trends Ecol. Evol. 10:236-240.

AUSIELLO, G., P. CRESCENZI, G. GAMBOSI, V. KANN, A. MARCHETTI-SPACCAMELA AND M. PROTASI. 1999. Complexity and approximation - combinatorial optimization problems and their approximability properties. Springer Verlag, Berlin.

BAHLO, M. AND R. C. GRIFFITHS. 2000. Inference from gene trees in a subdivided population. Theor. Popul. Biol. 57:79-95.

BAKKE, E. AND A. VON HAESELER. 1999. Distance measures in terms of substitution process. Theor. Popul. Biol. 55:166-175.

BANDELT, H. J. AND A. W. M. DRESS. 1992. A canonical decomposition-theory for metrics on a finite-set. Adv. Math. 92:47-105.

BARKER, N. P., H. P. LINDER AND E. H. HARLEY. 1995. Polyphyly of Arundinoideae (Poaceae): Evidence from *rbcL* sequence data. Syst. Bot. 20:423-435.

BARKER, N. P., H. P. LINDER AND E. H. HARLEY. 1999. Sequences of the grass-specific insert in the chloroplast *rpoC2* gene elucidate generic relationships of the Arundinoideae (Poaceae). Syst. Bot. 23:327-350.

BARRACLOUGH, T. G., J. E. HOGAN AND A. P. VOGLER. 1999. Testing whether ecological factors promote cladogenesis in a group of tiger beetles (Coleoptera : Cicindelidae). Proc. R. Soc. Lond. B 266:1061-1067.

BARRACLOUGH, T. G. AND S. NEE. 2001. Phylogenetics and speciation. Trends Ecol. Evol. 16:391-399.

BARRACLOUGH, T. G. AND A. P. VOGLER. 2000. Detecting the geographical pattern of speciation from species- level phylogenies. Am. Nat. 155:419-434.

BARRACLOUGH, T. G., A. P. VOGLER AND P. H. HARVEY. 1998. Revealing the factors that promote speciation. Phil. T. Roy. Soc. B 353:241-249.

BATES, D. M. AND D. G. WATTS. 1988. Nonlinear Regression Analysis and Its Applications. Wiley, New York.

BAUM, B. R. 1987. Numerical taxonomic analyses of the Poaceae *In* Grass systematics and evolution (T. R. Soderstrom, K. W. Hilu, C. S. Campbell and M. W. Barkworth, eds.). Pp. 334-342. Smithsonian Institution Press, Washington D.C.

BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of comibining gene tree. Taxon 41:3-10.

BAUM, B. R. AND M. A. RAGAN. 1993. Reply to A. G. Rodrigo' s "A comment on Baum' s method for combining phylogenetic trees". Taxon 42:637-640.

BAUM, D. A. AND A. LARSON. 1991. Adaptation reviewed - a phylogenetic methodology for studying character macroevolution. Syst. Zool. 40:1-18.

BEERLI, P. AND J. FELSENSTEIN. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. USA 98:4563-4568.

BENNETZEN, J. L. 2000. Comparative sequence analysis of plant nuclear genomes: microlinearity and its many exceptions. Plant Cell 12:1021-1030.

BENNETZEN, J. L. AND M. FREELING. 1993. Grasses as a single genetic system: genome composition, colinearity and compatibility. Trends Genet. 9:259-261.

BENTON, M. J. 1996. Testing the time axis of phylogenies *In* New uses for new phylogenies (P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith and S. Nee, eds.). Pp. 217-233. Oxford University Press, Oxford.

BERLOCHER, S. H. 1998. Can sympatric speciation be proven from phylogenetic and biogeographic evidence? *In* Endless forms: species and speciation (D. J. Howard and S. H. Berlocher, eds.). Pp. 99-113. Oxford University Press, Oxford, UK.

BERNER, R. A. AND Z. KOTHAVALA. 2001. GEOCARB III: A revised model of atmospheric CO2 over phanerozoic time. Am. J. Sci. 301:182-204.

BERRY, V. AND O. GASCUEL. 1996. On the interpretation of bootstrap trees: Appropriate threshold of clade selection and induced gain. Mol. Biol. Evol. 13:999-1011.

BESSEY, C. E. 1915. The phylogenetic taxonomy of flowering plants. Ann. Missouri Bot. Gard. 2:109-164.

BININDA-EMONDS, O. R. P. 2000. Factors influencing phylogenetic inference: A case study using the mammalian carnivores. Mol. Phylogenet. Evol. 16:113-126.

BININDA-EMONDS, O. R. P. AND H. N. BRYANT. 1998. Properties of matrix representation with parsimony analyses. Syst. Biol. 47:497-508.

BININDA-EMONDS, O. R. P., J. L. GITTLEMAN AND A. PURVIS. 1999. Building large trees by combining phylogenetic information: A complete phylogeny of the extant Carnivora (Mammalia). Biol. Rev. 74:143-175.

BININDA-EMONDS, O. R. P. AND M. J. SANDERSON. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. Syst. Biol. 50:565-579.

BISHOP, J. G., A. M. DEAN AND T. MITCHELL-OLDS. 2000. Rapid evolution in plant chitinases: Molecular targets of selection in plant-pathogen coevolution. Proc. Natl. Acad. Sci. USA 97:5322-5327.

BOCHERENS, H., P. L. KOCH, A. MARIOTTI, D. GERAADS AND J. J. JAEGER. 1996. Isotopic biogeochemistry (C-13, O-18) of mammalian enamel from African Pleistocene hominid sites. Palaios 11:306-318.

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.

BREMER, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. J. Herpetol. 42:795-803.

BROOKS, D. R. AND D. A. MCLENNAN. 1991. Phylogeny, ecology, and behavior: A research program in comparative biology. University of Chicago Press, Chicago.

BROWER, A. V. Z., R. DESALLE AND A. VOGLER. 1996. Gene trees, species trees, and systematics: A cladistic perspective. Annu. Rev. Ecol. Syst. 27:423-450.

BRYANT, D. AND M. STEEL. 1995. Extension operations on sets of leaf-labelled trees. Adv. Appl. Math. 16:425-453.

BUNEMAN, P. 1971. The recovery of trees from measures of dissimilarity *In* Mathematics in the archeological and historical sciences (F. R. Hodson, D. G. Kendall and P. Tautu, eds.). Pp. 387-395. Edinburgh University Press, Edinburgh.

BURT, A. 1989. Comparative methods using phylogenetically independent contrasts. Oxford University Press, Oxford.

BUSH, G. L. AND J. J. SMITH. 1998. The genetics and ecology of sympatric speciation: A case study. Res. Popul. Ecol. 40:175-187.

CAIN, A. J. 1959. The post-Linnean development of taxonomy. Proc. R. Soc. Lond. B 170:234-244.

CAMIN, J. H. AND R. R. SOKAL. 1965. A method for deducing branching sequences in phylogeny. Evolution 19:311-326.

CAMPBELL, C. S. AND E. A. KELLOGG. 1987. Sister group relationships of the Poaceae *In* Grass systematics and evolution (T. R. Soderstrom, K. W. Hilu, C. S. Campbell and M. E. Barkworth, eds.). Pp. 217-224. Smithsonian Institution Press, Washington D.C.

CAVALLI-SFORZA, L. L. AND A. W. F. EDWARDS. 1967. Phylogenetic analysis: models and estimation procedures. Evolution 21:550-570.

CERLING, T. E., J. M. HARRIS, B. J. MACFADDEN, M. G. LEAKEY, J. QUADE, V. EISENMANN AND J. R. EHLERINGER. 1997. Global vegetation change through the Miocene/Pliocene boundary. Nature 389:153-158.

CERLING, T. E., Y. WANG AND J. QUADE. 1993. Expansion of C_4 ecosystems as an indicator of global ecological change in the late Miocene. Nature 361:344-345.

CHAN, K. M. A. AND B. R. MOORE. 1999. Accounting for mode of speciation increases power and realism of tests of phylogenetic asymmetry. Am. Nat. 153:332-346.

CHANDLER, V. L. AND S. WESSLER. 2001. Grasses. A collective model genetic system. Plant Physiol. 125:1155-1156.

CHANG, J. T. 1996. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math. Biosci. 134:189-215.

CHAPMAN, G. P. 1996. The biology of grasses. CAB International, Oxon.

CHARLESTON, M. A. 2001. Hitch-hiking: A parallel heuristic search strategy, applied to the phylogeny problem. J. Comput. Biol. 8:79-91.

CHARLESWORTH, D., B. CHARLESWORTH AND G. A. T. MCVEAN. 2001. Genome sequences and evolutionary biology, a two-way interaction. Trends Ecol. Evol. 16:235-242.

CHASE, M. W. AND A. V. COX. 1998. Gene sequences, collaboration, and analysis of large data sets. Austr. Syst. Bot. 11:215-229.

CHASE, M. W., D. E. SOLTIS, R. G. OLMSTEAD, D. MORGAN, D. H. LES, B. D. MISHLER, M. R. DUVALL, R. A. PRICE, H. G. HILLS, Y.-L. QIU, K. A. KRON, J. H. RETTIG, E. CONTI, J. D. PALMER, J. R. MANHART, K. J. SYTSMA, H. J. MICHAEL, W. J. KRESS, K. G. KAROL, W. D. CLARK, M. HEDREN, B. S. GAUT, R. K. JANSEN, K. J. KIM, C. F. WIMPEE, J. F. SMITH, G. R. FURNIER, S. H. STRAUSS, Q. Y. XIANG, G. M. PLUNKETT, P. S. SOLTIS, S. M. SWENSEN, S. E. WILLIAMS, P. A. GADEK, C. J. QUINN, L. E. EGUIARTE, E. GOLENBERG, G. H. LEARN, S. W. GRAHAM, S. C. H. BARRETT, S. DAYANANDAN AND V. A. ALBERT. 1993. Phylogenetics of seed plants - an analysis of nucleotide-sequence from the plastid gene *rbcL*. Ann. Missouri Bot. Gard. 80:528-580.

CHAW, S. M., C. L. PARKINSON, Y. C. 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.

CHESSER, R. T. AND R. M. ZINK. 1994. Modes of speciation in birds - a test of Lynch' s method. Evolution 48:490-497.

CHEVERUD, J. M., M. M. DOW AND W. LEUTENEGGER. 1985. The quantitative assessment of phylogenetic constraints in comparative analyses: Sexual dimorphism in body weight among primates. Evolution 39:1335-1351.

CLARK, L. G., M. KOBAYASHI, S. MATHEWS, R. E. SPANGLER AND E. A. KELLOGG. 2000. The Puelioideae, a new subfamily of Poaceae. Syst. Bot. 25:181-187.

CLARK, L. G., W. ZHANG AND J. F. WENDEL. 1995. A phylogeny of the grass family (Poaceae) based on *ndhF* sequence data. Syst. Bot. 20:436-460.

CLAYTON, W. D. AND S. A. RENVOIZE. 1986. Genera Graminum, grass genera of the world. Her Majesty' s Stationery Office, London.

CLEGG, M. T., B. S. GAUT, G. H. LEARN AND B. R. MORTON. 1994. Rates and Patterns of Chloroplast Dna Evolution. Proc. Natl. Acad. Sci. USA 91: 6795-6801.

CLIFFORD, H. T., W. T. WILLIAMS AND G. N. LANCE. 1969. A further numerical contribution to the classification of the Poaceae. Austr. J. Bot. 17:119-131.

CONSTANTINESCU, M. AND D. SANKOFF. 1995. An efficient algorithm for supertrees. J. Classif. 12:101-112.

COYNE, J. A. AND T. D. PRICE. 2000. Little evidence for sympatric speciation in island birds. Evolution 54:2166-2171.

CREPET, W. L. AND G. D. FELDMANN. 1991. The earliest remains of grasses in the fossil record. Am. J. Bot. 78:1010-1014.

CRONQUIST, A. 1981. An integrated system of classification of flowering plants. Columbia University Press, New York.

CUMMINGS, M. P., L. M. KING AND E. A. KELLOG. 1994. Slipped-strand mispairing in a plastid gene: *rpoC2* in grasses (Poaceae). Mol. Biol. Evol. 11:1-8.

CUNNINGHAM, C. W., K. E. OMLAND AND T. H. OAKLEY. 1998. Reconstructing ancestral character states: a critical reappraisal. Trends Ecol. Evol. 13:361-366.

CUTLER, D. J. 2000. Estimating divergence times in the presence of an overdispersed molecular clock. Mol. Biol. Evol. 17:1647-1660.

DAGHLIAN, C. P. 1981. A review of the fossil record of monocotyledons. Bot. Rev. 47:517-555.

DAHLGREN, R., H. T. CLIFFORD AND P. F. YEO. 1985. The families of monocotyledons. Springer Verlag, Berlin.

DAVIS, J. I. AND R. J. SORENG. 1993. Phylogenetic structure in the grass family (Poaceae) as inferred from chloroplast DNA restriction site variation. Am. J. Bot. 80:1444-1454.

DE ROSA, R., J. K. GRENIER, T. ANDREEVA, C. E. COOK, A. ADOUTTE, M. AKAM, S. B. CARROLL AND G. BALAVOINE. 1999. Hox genes in brachiopods and priapulids and protostome evolution. Nature 399:772-776.

DEBRY, R. W. AND R. G. OLMSTEAD. 2000. A simulation study of reduced tree-search effort in bootstrap resampling analysis. Syst. Biol. 49:171-179.

DEBRY, R. W. AND N. A. SLADE, 1985. Cladistic analysis of restriction endonuclease cleavage maps within a maximum-likelihood framework. Syst. Zool. 34:21-34.

DENGLER, N. G. AND T. NELSON. 1999. Leaf structure and development in C_4 plants *In* C_4 Plant Biology (R. F. Sage and R. K. Monson, eds.). Pp. 133-172. Academic Press, San Diego.

DINIZ-FILHO, J. A. F. 2001. Phylogenetic autocorrelation under distinct evolutionary processes. Evolution 55:1104-1109.

DINIZ-FILHO, J. A. F., C. E. R. DE SANT'ANA AND L. M. BINI. 1998. An eigenvector method for estimating phylogenetic inertia. Evolution 52:1247-1262.

DOBZHANSKY, T. 1973. Nothing in biology makes sense except in the light of evolution. Am. Biol. Teach. 35:125-129.

DOEBLEY, J., M. DURBIN, E. M. GOLENBERG, M. T. CLEGG AND D. P. MA. 1990. Evolutionary analysis of the large subunit of carboxylase (*rbcL*) nucleotide sequence among the grasses (Gramineae). Evolution 44:1097-1108.

DONOGHUE, M. J., J. J. DOYLE, J. GAUTHIER, A. G. KLUGE AND T. ROWE. 1989. The importance of fossils in phylogeny reconstruction. Annu. Rev. Ecol. Syst. 20:431-460.

DOUST, A. N. AND E. A. KELLOGG. 2002. Inflorescence diversification in the panicoid 'bristle grass' clade (Paniceae, Poaceae): Evidence from molecular phylogenies and developmental morphology. Am. J. Bot. 89:1203-1222.

DOYLE, J. A., M. J. DONOGHUE AND E. A. ZIMMER. 1994. Integration of morphological and ribosomal-RNA data on the origin of angiosperms. Ann. Missouri Bot. Gard. 81:419-450.

DOYLE, J. J. 1997. Trees within trees: Genes and species, molecules and morphology. Syst. Biol. 46:537-553.

DOYLE, J. J., J. I. DAVIS, R. J. SORENG, D. GARVIN AND M. J. ANDERSON. 1992. Chloroplast DNA inversions and the origin of the grass family (Poaceae). Proc. Natl. Acad. Sci. USA 89:7722-7726.

Doyle, J. J. and J. L. Doyle. 1987. A rapid DNA isolation procedure for small amounts of fresh leaf tissue. Phytochem. Bull. 19: 11-15

DUGAS, D. P. AND G. J. RETALLACK. 1993. Middle miocene fossil grasses from Fort Ternan, Kenya. J. Paleontol. 67:113-128.

DUVALL, M. R. AND B. R. MORTON. 1996. Molecular phylogenetics of Poaceae: an expanded analysis of *rbcL* sequence data. Mol. Phylogenet. Evol. 5:353-358.

ECK, R. V. AND M. O. DAYHOFF. 1966. Atlas of protein sequence and structure. National Biomedical Research Foundation, Silver Spring.

EDWARDS, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge.

EDWARDS, A. W. F. AND L. L. CAVALLI-SFORZA. 1963. The reconstruction of evolution. Ann. Hum. Genet. 27:105.

EDWARDS, A. W. F. AND L. L. CAVALLI-SFORZA. 1964. Reconstruction of evolutionary trees *In* Phenetic and Phylogenetic Classification (V. H. Heywood and J. McNeill, eds.). Pp. 67-76. Systematics Association, London.

EFRON, B. 1979. Bootstrapping methods: another look at the jackknife. Ann. Stat. 7:1-26.

EFRON, B., E. HALLORAN AND S. HOLMES. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA 93:7085-7090.

EFRON, B. AND R. J. TIBSHIRANI. 1993. An Introduction to the bootstrap. Chapman and Hall, New York.

EMERSON, B. C., E. PARADIS AND C. THEBAUD. 2001. Revealing the demographic histories of species using DNA sequences. Trends Ecol. Evol. 16:707-716.

ERDÖS, P. L., M. A. STEEL, L. A. SZÉKELY AND T. J. WARNOW. 1997. Local quartet splits of a binary tree infer all quartet splits via one dyadic inference rule. Comput. Artif. Intell. 16:217-227.

FARRIS, J. S. 1970. Methods for computing Wagner trees. Syst. Zool. 19:83-92.

FARRIS, J. S. 1977. Phylogenetic analysis under Dollo' s Law. Syst. Zool. 26:77-88.

FARRIS, J. S., V. A. ALBERT, M. KALLERSJO, D. LIPSCOMB AND A. G. KLUGE. 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12:99-124.

FARRIS, J. S., M. KÄLLERSJÖ, A. G. KLUGE AND C. BULT. 1995. Constructing a significance test for incongruence. Syst. Biol. 44:570-572.

FARRIS, J. S., A. G. KLUGE AND M. M. ECKHARDT. 1970. A numerical approach to phylogenetic systematics. Syst. Zool. 19:172-191.

FELSENSTEIN, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Genet. 25:471-492.

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

FELSENSTEIN, J. 1978b. The number of evolutionary trees. Syst. Zool. 27:27-33.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368-376.

FELSENSTEIN, J. 1982. Numerical methods for inferring evolutionary trees. Quart. Rev. Biol. 379-404.

FELSENSTEIN, J. 1983. Parsimony in systematics: biological and statistical issues. Annu. Rev. Ecol. Syst. 14:313-333.

FELSENSTEIN, J. 1985a. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-789.

FELSENSTEIN, J. 1985b. Phylogenies and the comparative method. Am. Nat. 125:1-12.

FELSENSTEIN, J. 2000, 3.6. PHYLIP (Phylogeny Inference Package). Department of Genetics, University of Washington, Seattle

FELSENSTEIN, J. 2001. The troubled growth of statistical phylogenetics. Syst. Biol. 50:465-467.

FELSENSTEIN, J. AND H. KISHINO. 1993. Is there something wrong with the bootstrap on phylogenies - a reply. Syst. Biol. 42:193-200.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Syst. Biol. 20:406-416.

FITCH, W. M. 1974. Response to the paper of Dr. Moore *In* Genetic Distance (J. F. Crow and C. Dennison, eds.). Pp. 117-119. Plenum Press, New York.

FITCH, W. M. 1976. Molecular evolutionary clocks *In* Molecular evolution (F. J. Ayala, ed.). Pp. 160-179. Sinauer Associates, Sunderland.

FITCH, W. M. AND F. J. AYALA. 1994a. The superoxide-dismutase molecular clock revisited. Proc. Natl. Acad. Sci. USA 91:6802-6807.

FITCH, W. M. AND F. J. AYALA. 1994b. Tempo and mode in evolution. Proc. Natl. Acad. Sci. USA 91:6717-6720.

FITCH, W. M. AND J. S. FARRIS. 1974. Evolutionary trees with minimum nucleotide replacements from amino acid sequences. J. Mol. Evol. 3:263-278.

FITCH, W. M. AND E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science 155:279-284.

FITCH, W. M. AND E. MARKOVICH. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4:579-593.

FREELING, M. 2001. Grasses as a single genetic system. Reassessment 2001. Plant Physiol. 125:1191-1197.

FRIEDMAN, W. E. AND S. K. FLOYD. 2001. Perspective: The origin of flowering plants and their reproductive biology - A tale of two phylogenies. Evolution 55:217-231.

FU, Y. X. 1994. Estimating effective population size or mutation rate using the frequencies of mutations of various casses in a sample of DNA sequences. Genetics 138:1375-1386.

FUSCO, G. AND Q. C. B. CRONK. 1995. A new method for evaluating the shape of large phylogenies. J. Theor. Biol. 175:235-243.

GALTIER, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18:866-873.

GALTIER, N., N. TOURASSE AND M. GOUY. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science 283:220-221.

GAUT, B. S., L. G. CLARK, J. F. WENDEL AND S. V. MUSE. 1997. Comparisons of the molecular evolutionary process at *rbcL* and *ndhF* in the grass family (Poaceae). Mol. Biol. Evol. 14:769-777.

GAUT, B. S., S. V. MUSE, W. D. CLARK AND M. T. CLEGG. 1992. Relative rates of nucleotide substitution at the *rbcL* locus of monocotyledonous Plants. J. Mol. Evol. 35:292-303.

GAUT, B. S., A. S. PEEK, B. R. MORTON AND M. T. CLEGG. 1999. Patterns of genetic diversification within the *Adh* gene family in the grasses (Poaceae). Mol. Biol. Evol. 16:1086-1097.

GAUTHIER, J., A. G. KLUGE AND T. ROWE. 1988. Amniote phylogeny and the importance of fossils. Cladistics 12:152-162.

GEYER, C. J. 1991. Markov chain Monte Carlo maximum likelihood *In* Computing science and statistics: Proceedings of the 23rd symposium on the interface (W. M. Keramida, ed.). Pp. 156-163. Fairfax Station: Interface Foundation,

GIANNASI, N., A. MALHOTRA AND R. S. THORPE. 2001. Nuclear and mtDNA phylogenies of the *Trimeresurus* complex: Implications for the gene versus species tree debate. Mol. Phylogenet. Evol. 19:57-66.

GILLESPIE, J. H. AND C. H. LANGLEY. 1979. Are evolutionary rates really variable? J. Mol. Evol. 13:27-34.

GOLDING, G. B. AND A. M. DEAN. 1998. The structural basis of molecular adaptation. Mol. Biol. Evol. 15:355-369.

GOLDMAN, N. 1990. Maximum likelihood of phylogenetic trees, with special reference to Poisson process models of DNA substitution and to parsimony analysis. Syst. Zool. 39:345-361.

GOLDMAN, N., J. P. ANDERSON AND A. G. RODRIGO. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652-670.

GOLDSTEIN, D. B. AND P. H. HARVEY. 1999. Evolutionary inference from genomic data. Bioessays 21:148-156.

GORDON, A. D. 1980. On the assessment and comparison of classifications *In* Analyse de données et informatique (R. Tomassone, ed.). Pp. 149-160. INRA, LeChesnay.

GPWG. 2001. Phylogeny and subfamilial classification of the grasses (Poaceae). Ann. Missouri Bot. Gard. 88:373-457.

GRAFEN, A. 1989. The phylogenetic regression. Phil. T. Roy. Soc. B 326:119-157.

GRAFEN, A. AND M. RIDELY. 1996. Statistical tests for discrete cross-species data. J. Theor. Biol. 183:255-267.

GRAFEN, A. AND M. RIDLEY. 1997. A new model for discrete character evolution. J. Theor. Biol. 184:7-14.

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

GREEN, P. AND B. W. SILVERMAN. 1994. Nonparametric regression and generalized linear models: A roughness penalty approach. Chapman and Hall, New York.

GREEN, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711-732.

GU, X., Y. F. WANG AND J. Y. GU. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nature Genet. 31:205-209.

GUYER, C. AND J. B. SLOWINSKI. 1991. Comparisons of observed phylogenetic topologies with null expectations among three monophyletic lineages. Evolution 45:340-350.

HAMBY, R. K. AND E. A. ZIMMER. 1988. Ribosomal RNA sequences for inferring phylogeny within the grass family (Poaceae). Plant Syst. Evol. 160:29-37.

HANSEN, B. E. 1997. Approximate asymptotic *p*-values for structural-change tests. J. Bus. Econ. Stat. 15:60-67.

HARVEY, P. H., E. C. HOLMES, A. O. MOOERS AND S. NEE. 1994. Inferring evolutionary processes from molecular phylogenies *In* Models in phylogeny reconstruction (D. J. Siebert and D. J. Williams, eds.). Pp. 313-333. Systematics Association, Oxford.

HARVEY, P. H. AND S. NEE. 1996. What this book is about *In* New uses for new phylogenies (P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith and S. Nee, eds.). Pp. 1-11. Oxford University Press, Oxford.

HARVEY, P. H. AND M. D. PAGEL. 1991. The comparative method in evolutionary biology. Oxford University Press, London.

HASEGAWA, M. AND H. KISHINO. 1989. Heterogeneity of tempo and mode of mitochondrial DNA evolution among mammalian orders. Jap. J. Genet. 64:243-258.

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.

HASTINGS, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97-109.

HECKMAN, D. S., D. M. GEISER, B. R. EIDELL, R. L. STAUFFER, N. L. KARDOS AND S. B. HEDGES. 2001. Molecular evidence for the early colonization of land by fungi and plants. Science 293:1129-1133.

HEDGES, S. B. 1992. The number of replications needed for accurate estimation of the bootstrap *p*-value in phylogenetic studies. Mol. Biol. Evol. 9:366-369.

HEILBUTH, J. C. 2000. Lower species richness in dioecious clades. Am. Nat. 156:221-241.

HENDY, M. D. AND D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297-309.

HENNIG, W. 1950. Grundzuge einer Theorie der phylogenetischen Systematik. Deutscher Zentralverlag, Berlin.

HENNIG, W. 1966. Phylogenetic systematics. University of Illinois Press, Urbana.

HILLIS, D. M. 1996. Inferring complex phylogenies. Nature 383:130-131.

HILLIS, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47:1-8.

HILLIS, D. M. AND J. J. BULL. 1993. An empirical-test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182-192.

HILLIS, D. M. AND J. P. HUELSENBECK. 1994a. To tree the truth: Biological and numerical simulations of phylogeny *In* Molecular evolution of physiological processes (D. M. Fambrough, ed.). Pp. 55-67. Rockfeller University Press, New York.

HILLIS, D. M., J. P. HUELSENBECK AND C. W. CUNNINGHAM. 1994b. Application and accuracy of molecular phylogenies. Science 264:671-677.

HILLIS, D. M., J. P. HUELSENBECK AND D. L. SWOFFORD. 1994. Hobgoblin of phylogenetics. Nature 369:363-364.

HILLIS, D. M., B. K. MABLE AND C. MORITZ. 1996. Applications of molecular systematics: The state of the field and a look to the future *In* Molecular systematics (D. M. Hillis, C. Moritz and B. K. Mable, eds.). Pp. 515-543. Sinauer Associates, Sunderland.

HILU, K. W., L. A. ALICE AND H. P. LIANG. 1999. Phylogeny of Poaceae inferred from *matK* sequences. Ann. Missouri Bot. Gard. 86:835-851.

HODKINSON, T. R., M. W. CHASE, C. TAKAHASHI, I. J. LEITCH, M. D. BENNET AND S. A. RENVOIZE. 2002a. The use of DNA sequencing (ITS and *trnL-F*), AFLP and fluorescent in-situ hybridisation to study allopolyploid *Miscanthus* (Poaceae). Am. J. Bot. 89:279-286.

HODKINSON, T. R., M. W. CHASE, M. D. LLEDÓ, N. SALAMIN AND S. A. RENVOIZE. 2002b. Phylogenetics of *Miscanthus*, *Saccharum* and related genera (Saccharinae, Andropogoneae, Poaceae) based on DNA sequences from ITS nuclear ribosomal DNA and plastid *trnL* intron and *trnL-F* intergenic spacers. J. Plant Res. 115: 381-392.

HOLLAND, P. W. H. 1999. Gene duplication: Past, present and future. Semin. Cell Dev. Biol. 10:541-547.

HOUSWORTH, E. A. AND E. P. MARTINS. 2001. Random sampling of constrained phylogenies: Conducting phylogenetic analyses when the phylogeny is partially known. Syst. Biol. 50:628-639.

HSIAO, C., S. W. L. JACOBS, N. P. BARKER AND N. J. CHATTERTON. 1998. A molecular phylogeny of the subfamily Arundinoideae (Poaceae) based on sequences of rDNA. Austr. Syst. Bot. 11:41-52.

HSIAO, C., S. W. L. JACOBS, N. J. CHATTERTON AND K. H. ASAY. 1999. A molecular phylogeny of the grass family (Poaceae) based on the sequences of nuclear ribosomal DNA (ITS). Austr. Syst. Bot. 11:667-688.

HUELSENBECK, J. P. 1995. The robustness of two phylogenetic methods - four-taxon simulations reveal a slight superiority of maximum-likelihood over neighbor joining. Mol. Biol. Evol. 12:843-849.

HUELSENBECK, J. P. 1997. Is the Felsenstein zone a fly trap ? Syst. Biol. 46:69-74.

HUELSENBECK, J. P. 1998. Systematic bias in phylogenetic analysis: is the *Strepsiptera* problem solved? Syst. Biol. 47:519-537.

HUELSENBECK, J. P. 2002. Testing a covariotide model of DNA substitution. Mol. Biol. Evol. 19:698-707.

HUELSENBECK, J. P. AND J. P. BOLLBACK. 2001. Empirical and hierarchical Bayesian estimation of ancestral states. Syst. Biol. 50:351-366.

HUELSENBECK, J. P., D. M. HILLIS AND R. JONES. 1996. Parametric bootstrapping in molecular phylogenetics: applications and performance. *In* Molecular zoology: advances, stategies and protocols. (J. D. Ferraris and S. R. Palumbi, eds.). Pp. 19-45. Wiley-Liss, New York.

HUELSENBECK, J. P. AND F. RONQUIST. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755.

HUEY, R. B. AND A. F. BENNETT. 1987. Phylogenetic studies of coadaptation: Preferred temperatures versus optimal performance temperatures in lizards. Evolution 41:1098-1115.

HUGHES, A. L. 1999. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. J. Mol. Evol. 48:565-576.

IVICS, Z., P. B. HACKETT, R. H. PLASTERK AND Z. IZSVAK. 1997. Molecular reconstruction of sleeping beauty, a *Tc1*-like transposon from fish, and its transposition in human cells. Cell 91:501-510.

JACOBS, B. F., J. D. KINGSTON AND L. L. JACOBS. 1999. The origin of grass-dominated ecosystems. Ann. Missouri Bot. Gard. 86:590-643.

JANIS, C. M. 1993. Tertiary mammal evolution in the context of changing climates, vegetation, and tectonic events. Annu. Rev. Ecol. Syst. 24:467-500.

JANIS, C. M., J. DAMUTH AND J. M. THEODOR. 2000. Miocene ungulates and terrestrial primary productivity: Where have all the browsers gone? Proc. Natl. Acad. Sci. USA 97:7899-7904.

JERMANN, T. M., J. G. OPITZ, J. STACKHOUSE AND S. A. BENNER. 1995. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature 374:57-59.

JOHNSON, L. A. AND D. E. SOLTIS. 1998. Assessing congruences: empirical examples from molecular data *In* Molecular systematics of plants II. DNA sequencing (D. E. Soltis, P. S. Soltis and J. J. Doyle, eds.). Pp. 297-348. Kluwer Academic Publishers, Norwell.

JONES, E. S., N. L. MAHONEY, M. D. HAYWARD, I. P. ARMSTEAD, J. G. JONES, M. O. HUMPHREYS, I. P. KING, T. KISHIDA, T. YAMADA, F. BALFOURIER, G. CHARMET AND J. W. FORSTER. 2002. An enhanced molecular marker based genetic map of perennial ryegrass (*Lolium perenne*) reveals comparative relationships with other Poaceae genomes. Genome 45:282-295.

JUDD, W. S., C. S. CAMPBELL, E. A. KELLOGG AND P. F. STEVENS. 1999. Plant systematics: A phylogenetic approach. Sinauer Associates, Sunderland.

JUKES, T. H. AND C. R. CANTOR. 1969. Evolution of protein molecules *In* Mammalian protein metabolism (H. N. Munro, ed.). Pp. 21-132. Academic Press, New York.

KÄLLERSJÖ, M., J. S. FARRIS, M. W. CHASE, B. BREMER, M. F. FAY, C. J. HUMPHRIES, G. PETERSEN, O. SEBERG AND K. BREMER. 1998. Simultaneous parsimony jackknife analysis of 2538 *rbcL* DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. Plant Syst. Evol. 213:259-287.

KANAI, R. AND E. G. EDWARDS. 1999. The biochemistry of C_4 photosynthesis *In* C_4 Plant Biology (R. F. Sage and R. K. Monson, eds.). Pp. 49-87. Academic Press, San Diego.

KATAYAMA, H. AND Y. OGIHARA. 1996. Phylogenetic affinities of the grasses to other monocots as revealed by molecular analysis of chloroplast DNA. Curr. Genet. 29:572-581.

KELCHNER, S. A. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. Ann. Missouri Bot. Gard. 87:482-498.

KELLER, B. AND C. FEUILLET. 2000. Colinearity and gene density in grass genomes. Trends Plant Sci. 5:246251.

KELLOGG, E. A. 1998. Relationships of cereal crops and other grasses. Proc. Natl. Acad. Sci. USA 95:2005-2010.

KELLOGG, E. A. 1999. Phylogenetic aspects of the evolution of C_4 photosynthesis *In* C_4 Plant Biology (R. F. Sage and R. K. Monson, eds.). Pp. 411-443. Academic Press, San Diego.

KELLOGG, E. A. AND C. S. CAMPBELL. 1987. Phylogenetic analyses of the Gramineae *In* Grass systematics and evolution (T. R. Soderstrom, K. W. Hilu, C. S. Campbell and M. E. Barkworth, eds.). Pp. 310-322. Smithsonian Institution Press, Washington D.C.

KELLOGG, E. A. AND H. P. LINDER. 1995. Phylogeny of Poales *In* Monocotyledons: Systematics and evolution (P. J. Rudall, P. J. Cribb, D. F. Cutler and C. J. Humphries, eds.). Pp. 511-542. Royal Botanic Gardens, Kew.

KIM, J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Syst. Biol. 45:363-374.

KIM, J. 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. Syst. Biol. 47:43-60.

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

KINGMAN, J. F. C. 1982. On the genealogy of large populations. J. Appl. Prob. 19:27-43.

KISHINO, H. AND M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170-179.

KLUGE, A. G. AND J. S. FARRIS. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18:1-32.

KUHNER, M. K. AND J. FELSENSTEIN. 1994. Simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459-468.

KUHNER, M. K., J. YAMATO AND J. FELSENSTEIN. 1995. Estimating effective population size and mutation rate fromsequence data using Metropolis-Hastings sampling. Genetics 140:1421-1430.

KUHNER, M. K., J. YAMATO AND J. FELSENSTEIN. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149:429-434.

KUMAR, S. AND S. B. HEDGES. 1998. A molecular timescale for vertebrate evolution. Nature 392:917-920.

LAPOINTE, F.-J. AND G. CUCUMEL. 1997. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. Syst. Biol. 46:306-312.

LARCHER, W. 1995. Physiological plant ecology: ecophysiology of functional groups. Springer Verlag, Berlin.

LARGET, B. AND D. L. SIMON. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750-759.

LAUDER, G. V., A. M. LEROI AND M. R. ROSE. 1993. Adaptations and History. Trends Ecol. Evol. 8:294-297.

LEGENDRE, P. AND V. MAKARENKOV. 2002. Reconstruction of biogeographic and evolutionary networks using reticulograms. Syst. Biol. 51:199-216.

LEHMANN, E. L. 1983. Theory of point estimation. Wiley, New York.

LEOPOLD, E. B. AND M. F. DENTON. 1987. Comparative age of grassland and steppe east and west of the northern Rocky Mountains. Ann. Missouri Bot. Gard. 74:841-867.

LEOPOLD, E. B., G. LIU AND S. CLAY-POOLE. 1992. Low-biomass vegetation in the Oligocene? *In* Eocene-Oligocene climatic and biotic evolution (D. R. Prothero and W. A. Berggren, eds.). Pp. 399-420. Princeton University Press, New Jersey.

LEWIS, P. O. 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. Mol. Biol. Evol. 15:277-283.

LEWIS, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913-925.

LI, P. AND J. BOUSQUET. 1992. Relative-rate test for nucleotide substitutions between two lineages. Mol. Biol. Evol. 9:1185-1189.

LI, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland.

LI, W.-H. AND D. GRAUR. 1991. Fundamentals of molecular evolution. Sinauer Associates, Sunderland.

LIANG, H. AND K. W. HILU. 1995. Application of the *matK* gene sequences to grass systematics. Can. J. Bot. 74:125-134.

LINDER, H. P. 1987. The evolutionary history of the Poales/Restionales - A hypothesis. Kew Bull. 42:297-318.

LIPSCOMB, D. L., J. S. FARRIS, M. KALLERSJO AND A. TEHLER. 1998. Support, ribosomal sequences and the phylogeny of the eukaryotes. Cladistics 14:303-338.

LOCKHART, P. J., A. W. D. LARKUM, M. A. STEEL, P. J. WADDELL AND D. PENNY. 1996. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA 93:1930-1934.

LOCKHART, P. J., M. A. STEEL, A. C. BARBROOK, D. H. HUSON, M. A. CHARLESTON AND C. J. HOWE. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. 15:1183-1188.

LOCKHART, P. J., M. A. STEEL, M. D. HENDY AND D. PENNY. 1994. Recovering evolutionary trees under more realistic model of sequence evolution. Mol. Biol. Evol. 42:605-612.

LONG, S. P., 1999. Environmental responses *In* C_4 Plant Biology (R. F. Sage and R. K. Monson, eds). Pp. 215-249. Academic Press, San Diego.

LOPEZ, P., D. CASANE AND H. PHILIPPE. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19:1-7.

LYNCH, M. 1990. Methods for the analysis of comparative data in evolutionary biology. Evolution 45:1065-1080.

LYNCH, M. AND P. E. JARRELL. 1993. A method for calibrating molecular clocks and its application to animal mitochondrial-DNA. Genetics 135:1197-1208.

LYONS-WEILER, J. AND G. A. HOELZER. 1997. Escaping from the Felsenstein zone by detecting long branches in phylogenetic data. Mol. Phylogenet. Evol. 8:375-384.

MABBERLEY, D. J. 1993. The plant-book: A portable dictionary of the vascular plants. 2nd. Cambridge University Press, Cambridge.

MACFADDEN, B. J. 1998. Tale of two rhinos: isotopic ecology, paleodiet, and niche differentiation of Aphelops and Teleoceras from the Florida Neogene. Paleobiology 24:274-286.

MACFADDEN, B. J. AND T. E. CERLING. 1994. Fossil horses, carbon isotopes and global change. Trends Ecol. Evol. 9:481-486.

McLysaght, A., K. Hokamp and K. H. Wolfe. 2002. Extensive genomic duplication during early chordate evolution. Nature Genet. 31:200-204.

Maddala, G. S. and I.-M. Kim. 1998. Unit roots, cointegration, and structural change. Cambridge University Press, Cambridge.

Maddison, D. R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. Syst. Zool. 40:315-328.

Maddison, D. R., M. D. Baker and K. A. Ober. 1999. Phylogeny of carabid beetles inferred from 18S ribosomal DNA (Coleoptera: Carabidae). Syst. Entomol. 24:103-138.

Maddison, W. P. 1989. Reconstructing character evolution on polytomous cladograms. Cladistics 5:365-377.

Maddison, W. P. 1990. A method for testing the correlated evolution of two binary characters: Are gains or losses concentrated on certain branches of a phylogenetic tree? Evolution 44:539-557.

Maddison, W. P. 1991. Squared-change parsimony reconstructions of ancestral states for continuous valued characters on a phylogenetic tree. Syst. Zool. 40:304-314.

Maddison, W. P. 1995. Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. Syst. Biol. 44:474-481.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523-536.

Makarenkov, V. and P. Legendre. 2001. Optimal variable weighting for ultrametric and additive trees and K-means partitioning: Methods and software. J. Classif. 18:245-271.

Martin, W., D. Lydiate, H. Brinkmann, G. Forkmann, H. Saedler and R. Cerff. 1993. Molecular phylogenies in angiosperm evolution. Mol. Biol. Evol. 10:140-162.

Martins, E. P. and T. F. Hansen. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am. Nat. 149:646-667.

Mason-Gamer, R. J., C. F. Weil and E. A. Kellogg. 1998. Granule-bound starch synthase: structure, function, and phylogenetic utility. Mol. Biol. Evol. 15:1658-1673.

Mathews, S. and R. A. Sharrock. 1996. The phytochrome gene family in grasses (Poaceae): a phylogeny and evidence that grasses have a subset of the loci found in dicot Angiosperms. Mol. Biol. Evol. 13:1141-1150.

Mathews, S., R. C. Tsai and E. A. Kellogg. 2000. Phylogenetic structure in the grass family (Poaceae): Evidence from the nuclear gene phytochrome B. Am. J. Bot. 87:96-107.

Mau, B. and M. A. Newton. 1997. Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. J. Comput. Graph. Stat. 6:122-131.

Maxam, A. M. and W. Gilbert. 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA 74:560-564.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21:1087-1091.

MIADLIKOWSKA, J. AND F. LUTZONI. 2000. Phylogenetic revision of the genus *Peltigera* (lichen-forming Ascomycota) based on morphological, chemical and large subunit nuclear ribosomal DNA data. Int. J. Plant Sci. 161:925-958.

MICHELMORE, R. W. AND B. C. MEYERS. 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res. 8:1113-1130.

MIYAMOTO, M. M. AND W. M. FITCH. 1995. Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol. 12:503-513.

MOOERS, A. O. AND D. SCHLUTER. 1999. Reconstructing ancestor states with maximum likelihood: Support for one- and two-rate models. Syst. Biol. 48:623-633.

MOORE, G. W. 1974. A counterexample to Fitch' s method for maximum parsimony trees*In* Genetic distance (J. F. Crow and C. Denniston, eds.). Pp. 105-116. Plenum Press, New York.

MOORE, G. W. 1977. Proof of the populous path algorithm for missing mutations in parsimonious trees. J. Theor. Biol. 66:95-106.

MOORE, G. W., M. GOODMAN AND J. BARNABAS. 1973. An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets. J. Theor. Biol. 38:423-457.

MORITZ, C. AND D. M. HILLIS. 1996. Molecular systematics: Context and controversies *In* Molecular systematics (D. M. Hillis, C. Moritz and B. K. Mable, eds.). Pp. 1-17. Sinauer Associates, Sunderland.

MORT, M. E., P. S. SOLTIS, D. E. SOLTIS AND M. L. MABRY. 2000. Comparison of three methods for estimating internal support on phylogenetic trees. Syst. Biol. 49:160-171.

MOUNT, D. W. 2001. Bioinformatics: Sequence and genome analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

MUELLER, L. D. AND F. J. AYALA. 1982. Estimation and interpretation of genetic distance in empirical studies. Genet. Res. 40:127-137.

MULLIS, K. B. AND F. A. FALOONA. 1987. Specific synthesis of DNA in vitro via a polymerase catalyzed chain reaction. Meth. Enzymol. 155:335-350.

MUSE, S. V. AND B. S. WEIR. 1992. Testing for equality of evolutionary rates. Genetics 132:269-276.

NADOT, S., R. BAJON AND B. LEJEUNE. 1994. The chloroplast gene *rps4* as a tool for the study of Poaceae phylogeny. Plant Syst. Evol. 191:27-38.

NANNEY, D. L., R. M. PREPARATA, F. P. PREPARATA, E. B. MEYER AND E. M. SIMON. 1989. Shifting ditypic site analysis: Heuristics for expanding the phylogenetic range of nucleotide sequences in Sankoff analysis. J. Mol. Evol. 28:451-459.

NEE, S. 2001. Inferring speciation rates from phylogenies. Evolution 55:661-668.

NEE, S., E. C. HOMES, R. M. MAY AND P. H. HARVEY. 1994. Extinction rates can be estimated from molecular phylogenies. Philos. Trans. R. Soc. Lond., Ser. B 349:25-31.

NEI, M., X. GU AND T. SITNIKOVA. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. USA 94:7799-7806.

NIELSEN, R. AND Z. H. YANG. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

NIXON, K. C. 1999. The Parsimony Ratchet, a new method for rapid parsimony analysis. Cladistics 15:407-414.

OGIHARA, Y., K. ISONO, T. KOJIMA, A. ENDO, M. HANAOKA, T. SHIINA, T. TERACHI, S. UTSUGI, M. MURATA, N. MORI, S. TAKUMI, K. IKEO, T. GOJOBORI, R. MURAI, K. MURAI, Y. MATSUOKA, Y. OHNISHI, H. TAJIRI AND K. TSUNEWAKI. 2002. Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. Mol. Genet. Genomics 266:740-746.

OLSEN, G. K., H. MATSUDA, R. HAGSTROM AND R. OVERBEEK. 1994. fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. CABIOS 10:41-48.

OMILIAN, A. R. AND D. J. TAYLOR. 2001. Rate acceleration and long-branch attraction in a conserved gene of cryptic Daphniid (Crustaceae) species. Mol. Biol. Evol. 18:2201-2212.

PAGANI, M., K. H. FREEMAN AND M. A. ARTHUR. 1999. Late Miocene atmospheric CO_2 concentrations and the expansion of C_4 grasses. Science 285:876-879.

PAGE, R. D. M. AND M. A. CHARLESTON. 1997. From gene to organismal phylogeny: Reconciled trees and the gene tree species tree problem. Mol. Phylogenet. Evol. 7:231-240.

PAGE, R. D. M. AND E. C. HOLMES. 1998. Molecular evolution: A phylogenetic approach. Blackwell Science, London.

PAGEL, M. 1994. Detecting correlated evolution on phylogenies - a general method for the comparative analysis of discrete characters. Proc. R. Soc. Lond. B 255:37-45.

PAGEL, M. 1997. Inferring evolutionary processes from phylogenies. Zool. Scripta 26:331-348.

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

PANCHEN, A. L. 1992. Classification, evolution, and the nature of biology. Cambridge University Press, Cambridge.

PELLMYR, O. AND J. LEEBENS-MACK. 1999. Forty million years of mutualism: evidence for Eocene origin of the yucca-yucca moth association. Proc. Natl. Acad. Sci. USA 96:9178-9183.

PENNY, D. AND M. D. HENDY. 1985. Testing methods of evolutionary tree construction. Cladistics 1:266-278.

PENNY, D., M. D. HENDY AND M. A. STEEL. 1992. Progress with methods for constructing evolutionary trees. Trends Ecol. Evol. 7:73-79.

PENNY, M. AND M. HASEGAWA. 1997. The platypus put in its place. Nature 387:549-550.

POSADA, D. AND K. A. CRANDALL. 1998. ModelTest: testing the model of DNA substitution. Bioinformatics 14:817-818.

POSTLETHWAIT, J. H., I. G. WOODS, P. NGO-HAZELETT, Y. L. YAN, P. D. KELLY, F. CHU, H. HUANG, A. HILL-FORCE AND W. S. TALBOT. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. Genome Res. 10:1890-1902.

PURVIS, A. 1995. A modification to Baum and Ragan' s method for combining phylogenetic trees. Syst. Biol. 44:251-255.

PURVIS, A. 1996. Using interspecies phylogenies to test macroevolutionary hypotheses *In* New uses for new phylogenies (P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith and S. Nee, eds.). Pp. 153-168. Oxford University Press, Oxford.

PURVIS, A., J. L. GITTLEMAN AND H.-K. LUH. 1994. Truth or consequences: effects of phylogenetic accuracy on two comparative method. J. Theor. Biol. 167:293-300.

PURVIS, A., S. NEE AND P. H. HARVEY. 1995. Macroevolutionary inferences from primate phylogeny. Proc. R. Soc. London B 260:329-333.

PURVIS, A. AND D. L. J. QUICKE. 1997. Building phylogenies: Are the big easy? Trends Ecol. Evol. 12:49-50.

PYBUS, O. G. AND P. H. HARVEY. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. Lond. B 267:2267-2272.

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? Mol. Biol. Evol. 18:1745-1753.

QIU, Y.-L., J. LEE, F. BERNASCONI-QUADRONI, D. E. SOLTIS, P. S. SOLTIS, M. ZANIS, Z. CHEN, V. SAVOLAINEN AND M. W. CHASE. 2000. Phylogeny of basal angiosperms: analysis of five genes from three genomes. Int. J. Plant Sci. 161:S3-S27.

QUICKE, D. L. J., J. TAYLOR AND A. PURVIS. 2001. Changing the landscape: A new strategy for estimating large phylogenies. Syst. Biol. 50:60-66.

RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phylogenet. Evol. 1:53-58.

RANNALA, B., J. P. HUELSENBECK, Z. H. YANG AND R. NIELSEN. 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47:702-710.

RANNALA, B. AND Z. H. YANG. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304-311.

RAUP, D. M., S. J. GOULD, T. J. M. SCHOPF AND D. S. SIMBERLOFF. 1973. Stochastic models of phylogeny and the evolution of diversity. J. Geol. 81:525-542.

RAVEN, P. H., R. H. EVERT AND S. E. EICHHORN. 1992. Biology of plants. 5th Edition. Worth Publishers, New York.

REEVE, H. K. AND P. W. SHERMAN. 1993. Adaptation and the goals of evolutionary research. Quaterly Review in Biology 68:1-32.

RENVOIZE, S. AND W. CLAYTON. 1992. Classification and evolution of the grasses *In* Grass evolution and domestication (G. Chapman, ed.). Pp. 3-37. Cambridge University Press, Cambridge.

RICE, K. A., M. J. DONOGHUE AND R. G. OLMSTEAD. 1997. Analyzing large data sets: rbcL 500 revisited. Syst. Biol. 46:554-563.

RIDLEY, M. 1983. The explanation of organic diversity: The comparative method and adaptations of mating. Clarendon Press, Oxford.

ROBINSON, D. F. AND L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci. 59:131-144.

RODRIGUEZ, F. J., J. L. OLIVER, A. MARIN AND J. R. MEDINA. 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. 142:485-501.

RODRIGUEZ-TRELLES, F., R. TARRIO AND F. J. AYALA. 2002. A methodological bias toward overestimation of molecular evolutionary time scales. Proc. Natl. Acad. Sci. USA 99:8112-8115.

RODRIGUO, A. G. 1993. Calibrating the bootstrap test of monophyly. Int. J. Parasitol. 23:507-514.

ROGERS, J. S. 1986. Deriving phylogenetic trees from allele frequencies. Syst. Zool. 35:297-310.

ROGERS, J. S. AND D. L. SWOFFORD. 1998. A fast method for approximating maximum likelihoods of phylogenetic trees from nucleotide sequences. Syst. Biol. 47:77-89.

ROHLF, F. J. 2001. Comparative methods for the analysis of continuous variables: Geometric interpretations. Evolution 55:2143-2160.

RONQUIST, F. 1996. Matrix representation of trees, redundancy, and weighting. Syst. Biol. 45:247-253.

RONQUIST, F. 1998. Fast Fitch-parsimony algorithms for large data sets. Cladistics 14:387-400.

ROYER, D. L., S. L. WING, D. J. BEERLING, D. W. JOLLEY, P. L. KOCH, L. J. HICKEY AND R. A. BERNER. 2001. Paleobotanical evidence for near present-day levels of atmospheric CO_2 during part of the tertiary. Science 292:2310-2313.

RUVINSKY, I., L. M. SILVER AND J. J. GIBSON-BROWN. 2000. Phylogenetic analysis of T-Box genes demonstrates the importance of amphioxus for understanding evolution of the vertebrate genome. Genetics 156:1249-1257.

RYDIN, C., M. KÄLLERSJÖ AND E. M. FRIIST. 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.

SALAMIN, N., T. R. HODKINSON AND V. SAVOLAINEN. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). Syst. Biol. 51:136-150.

SALAMIN, N., M. W. CHASE, T. R. HODKINSON AND V. SAVOLAINEN. in press. Assessing internal support with large phylogenetic DNA matrices. Mol. Phylogenet. Evol.

SALTER, L. A. AND D. K. PEARL. 2001. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. Syst. Biol. 50:7-17.

SANDERSON, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14:1218-1231.

SANDERSON, M. J. 2002a. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Mol. Biol. Evol. 19:101-109.

SANDERSON, M. J. 2002b. r8s, version 1.5. Section of Evolution and Ecology, University of California, Davis

SANDERSON, M. J. AND J. A. DOYLE. 2001. Sources of error and confidence intervals in estimating the age of angiosperms from *rbcL* and 18S rDNA data. Am. J. Bot. 88:1499-1516.

SANDERSON, M. J., A. PURVIS AND C. HENZE. 1998. Phylogenetic supertrees: assembling the trees of life. Trends Ecol. Evol. 13:105-109.

SANDERSON, M. J. AND M. F. WOJCIECHOWSKI. 2000. Improved bootstrap confidence limits in large-scale phylogenies, with an example from *Neo-Astragalus* (leguminosae). Syst. Biol. 49:671-685.

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.

SANGER, F., S. NICKLEN AND A. R. COULSON. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74:5463-5467.

SANKOFF, D. 1975. Minimal mutation trees of sequences. SIAM J. Appl. Math. 28:35-42.

SANKOFF, D. AND P. ROUSSEAU. 1975. Locating the vertices of a Steiner tree in arbitrary space. Math. Prog. 9:240-246.

SATO, A., C. O' HUIGIN, F. FIGUEROA, P. R. GRANT, B. R. GRANT, H. TICHY AND J. KLEIN. 1999. Phylogeny of Darwin' s finches as revealed bymtDNA sequences. Proc. Natl. Acad. Sci. USA 96:5101-5106.

SAVOLAINEN, V., M. W. CHASE, C. M. MORTON, S. B. HOOT, D. E. SOLTIS, C. BAYER, M. F. FAY, A. DEBRUIJN, S. SULLIVAN AND Y.-L. QIU. 2000. Phylogenetics of flowering plants based upon a combined analysis of plastid *atpB* and *rbcL* gene sequences. Syst. Biol. 49:306-362.

SAVOLAINEN, V., M. W. CHASE, N. SALAMIN, D. E. SOLTIS, P. S. SOLTIS, A. LOPEZ, O. FÉDRIGO AND G. J. P. NAYLOR. 2002. Phylogeny reconstruction and functional constraints in organellar genomes: Plastid versus animal Mitochondrion. Syst. Biol. 51:638-647.

SCHAAL, B. A. AND W. J. LEVERICH. 2001. Plant population biology and systematics. Taxon 50:679-695.

SCHLUTER, D. 2001. Ecology and the origin of species. Trends Ecol. Evol. 16:372-380.

SCHLUTER, D., T. PRICE, A. O. MOOERS AND D. LUDWIG. 1997. Likelihood of ancestor states in adaptive radiation. Evolution 51:1699-1711.

SEMPLE, C. AND M. STEEL. 2000. A supertree method for rooted trees. Discret Appl. Math. 105:147-158.

SEMPLE, C. AND K. H. WOLFE. 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. J. Mol. Evol. 48:555-564.

SHIMODAIRA, H. AND M. HASEGAWA. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114-1116.

SIDDALL, M. E. 1998. Succes of parsimony in the four-taxon case: long branch repulsion by likelihood in the Farris zone. Cladistics 14:209-220.

SIMMONS, M. P., H. OCHOTERENA AND T. G. CARR. 2001. Incorporation, relative homoplasy, and effect of gap characters in sequence-based phylogenetic analyses. Syst. Biol. 50:454-462.

SINGH, V. 1988. Hydrologic systems. Prentice Hall, New Jersey.

SLOWINSKI, J. B. AND C. GUYER. 1989. Testing the stochasticity of patterns of organismal diversity: an improved null model. Am. Nat. 134:907-921.

SMALL, R. L. AND J. F. WENDEL. 2000. Phylogeny, duplication, and intraspecific variation of *Adh* sequences in new world diploid cottons (*Gossypium* L., Malvaceae). Mol. Phylogenet. Evol. 16:73-84.

SMITH, J. F. 2001. High species diversity in fleshy-fruited tropical understory plants. Am. Nat. 157:646-653.

SOKAL, R. R. AND F. J. ROHLF. 1995. Biometry: the principles and practice of statistics in biological research. 3rd Edition. W. H. Freeman, New York.

SOKAL, R. R. AND P. H. A. SNEATH. 1963. Principles of Numerical Taxonomy. W. H. Freeman, San Francisco.

SOLTIS, D. E., P. S. SOLTIS, V. A. ALBERT, D. G. OPPENHEIMER, C. W. DEPAMPHILIS, H. MA, M. W. FROHLICH AND G. THEISSEN. 2002a. Missing links: the genetic architecture of flower and floral diversification. Trends Plant Sci. 7:22-31.

SOLTIS, P. S., D. E. SOLTIS, V. SAVOLAINEN, P. R. CRANE AND T. G. BARRACLOUGH. 2002b. Rate heterogeneity among lineages of tracheophytes: Integration of molecular and fossil data and evidence for molecular living fossils. Proc. Natl. Acad. Sci. USA 99:4430-4435.

SOLTIS, D. E., P. S. SOLTIS, M. W. CHASE, M. E. MORT, D. C. ALBACH, M. ZANIS, V. SAVOLAINEN, W. H. HAHN, S. B. HOOT, M. F. FAY, M. AXTELL, S. M. SWENSEN, L. M. PRINCE, W. J. KRESS, K. C. NIXON AND J. S. FARRIS. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. Bot. J. Linnean Soc. 133:381-461.

SOLTIS, D. E., P. S. SOLTIS, M. E. MORT, M. W. CHASE, V. SAVOLAINEN, S. B. HOOT AND C. M. MORTON. 1998. Inferring complex phylogenies using parsimony: an empirical approach using three large DNA data sets for angiosperms. Syst. Biol. 47:32-42.

SOLTIS, D. E., P. S. SOLTIS, D. L. NICKRENT, L. A. JOHNSON, W. J. 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. Missouri Bot. Gard. 84:1-49.

SOLTIS, P. S. AND D. E. SOLTIS. 2001. Molecular systematics: assembling and using the Tree of Life. Taxon 50:663-677.

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.

SORENG, R. J. AND J. I. DAVIS. 1998. Phylogenetics and character evolution in the grass family (Poaceae): simultaneous analysis of morphological and chloroplast DNA restriction site character sets. Bot. Rev. 64:1-85.

SORENG, R. J., J. I. DAVIS AND J. J. DOYLE. 1990. A phylogenetic analysis of chloroplast DNA restriction site variation in Poaceae subfam. *Pooideae*. Plant Syst. Evol. 172:83-97.

SPORNE, K. R. 1956. The Phylogenetic classification of the Angiosperms. Cambridge University Press, Cambridge.

SPORNE, K. R. 1974. The morphology of Gymnosperms. 2nd edition. Hutchinson, London.

STEBBINS, G. L. 1956. Cytogenetics and evolution of the grass family. Am. J. Bot. 43:890-905.

STEEL, M. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. J. Classif. 9:91-116.

STEEL, M. 2001. Sufficient conditions for two tree reconstruction techniques to succeed on sufficiently long sequences. SIAM J. Discret. Math. 14:36-48.

STEEL, M., A. W. M. DRESS AND S. BOCKER. 2000. Simple but fundamental limitations on supertree and consensus tree methods. Syst. Biol. 49:363-368.

STEEL, M. AND A. MCKENZIE. 2001. Properties of phylogenetic trees generated by Yule-type speciation models. Math. Biosci. 170:91-112.

STEEL, M. AND D. PENNY. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17:839-850.

STEEL, M. A., L. SZÉKELY, P. L. ERDÖS AND P. J. WADDELL. 1993. A complete family of phylogenetic invariants for any number of taxa under Kimura' s 3ST model. N. Z. J. Bot. 31:289-296.

STEEL, M. A., L. SZÉKELY AND M. D. HENDY. 1994. Reconstructing trees from sequences whose sites evolve at variable rates. J. Comput. Biol. 1:153-163.

STRIMMER, K. AND V. MOULTON. 2000. Likelihood analysis of phylogenetic networks using directed graphical models. Mol. Biol. Evol. 17:875-881.

STRIMMER, K. AND A. VON HAESELER. 1996. Accuracy of neighbor joining for n-taxon trees. Syst. Biol. 45:516-523.

STRIMMER, K., C. WIUF AND V. MOULTON. 2001. Recombination analysis using directed graphical models. Mol. Biol. Evol. 18:97-99.

SWOFFORD, D. L. 2000, PAUP*4. Phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland

SWOFFORD, D. L. AND S. H. BERLOCHER. 1987. Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. Syst. Zool. 36:293-325.

SWOFFORD, D. L. AND W. P. MADDISON. 1987. Reconstructing ancestral character states under Wagner parsimony. Math. Biosci. 87:199-229.

SWOFFORD, D. L., G. K. OLSEN, P. J. WADDELL AND D. M. HILLIS. 1996. Phylogeny reconstruction *In* Molecular systematics (D. M. Hillis, C. Moritz and B. K. Mable, eds.). Pp. 407-514. Sinauer Associates, Sunderland.

SWOFFORD, D. L., P. J. WADDELL, J. P. HUELSENBECK, P. G. FOSTER, P. O. LEWIS AND J. S. ROGERS. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525-539.

TABERLET, P., L. GIELLY, G. PAUTOU AND J. BOUVET. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Mol.Biol. 17:1105-1109.

TAJIMA, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135: 599-607.

TAKEZAKI, N., A. RZHETSKY AND M. NEI. 1995. Phylogenetic test of the molecular clock and linearized trees. Mol. Biol. Evol. 12:823-833.

TAMURA, K. AND M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512-526.

THIELE, K. 1993. The holy grail of the perfect character: the cladistic treatment of morphometric data. Cladistics 9: 275-304.

THOMASSON, J. R. 1987. Fossil grasses: 1820-1986 and beyond *In* Grass systematics and evolution (T. R. Soderstrom, K. W. Hilu, C. S. Campbell and M. E. Barkworth, eds.). Pp. 159-167. Smithsonian Institution Press, Washington D.C.

THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN AND D. G. HIGGINS. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876-4882.

THORNE, J. L., H. KISHINO AND I. S. PAINTER. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647-1657.

TIERNEY, L. 1994. Markov chains for exploring posterior distributions. Ann. Statist. 22: 1701-1728.

TUFFLEY, C. AND M. STEEL. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59:581-607.

UYENOYAMA, M. K. 1995. A generalized least-squares estimate for the origin of sporophytic self-incompatibility. Genetics 139:975-992.

VON HAESELER, A. AND M. SCHONIGER. 1995. Modeling evolution of autocorrelated sequences. Comput. Stat. Data Anal. 20:457-458.

WATSON, L. A. AND M. J. DALLWITZ. 1992. The grass genera of the world. CAB International, Wallingford.

WEBB, S. D., R. C. J. HULBERT AND W. D. LAMBERT. 1995. Climatic implications of large herbivore distributions in the Miocene of North America *In* Paleo-climate and evolution, with emphasis on human origins (E. S. Vrba, G. H. Denton, T. C. Partridge and L. H. Burckle, eds.). Pp. 91-108. Yale University Press, New Haven.

WENDEL, J. F. AND J. J. DOYLE. 1998. Phylogenetic incongruence: window into genome history and molecular evolution *In* Molecular systematics of plants II. DNA sequencing (D. E. Soltis, P. S. Soltis and J. J. Doyle, eds.). Pp. 264-296. Kluwer Academic Publishers, Norwell.

WHEELER, W. C. AND K. NIXON. 1995. A novel method for economical diagnosis of cladograms under Sankoff optimization. Cladistics 10:207-213.

WHITTINGHAM, L. A., B. SLIKAS, D. W. WINKLER AND F. H. SHELDON. 2002. Phylogeny of the tree swallow genus, *Tachycineta* (Aves : Hirundinidae), by Bayesian analysis of mitochondrial DNA sequences. Mol. Phylogenet. Evol. 22:430-441.

WIENS, J. J. 1995. Polymorphic characters in phylogenetic systematics. Syst. Biol. 44:482-500.

WIENS, J. J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst. Biol. 47:625-640.

WIENS, J. J. 2001. Character analysis in morphological phylogenetics: Problems and solutions. Syst. Biol. 50:689-699.

WILKINSON, M. 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Syst. Biol. 44:501-514.

WILKINSON, M., J. L. THORLEY AND P. UPCHURCH. 2000. A chain is no stronger than its weakest link: Double decay analysis of phylogenetic hypotheses. Syst. Biol. 49:754-776.

WILLIS, K. AND J. MCELWAIN. 2002. The evolution of plants. Oxford University Press, Oxford.

WILSON, A. C., S. S. CARLSON AND T. J. WHITE. 1977. Biochemical evolution. Annu. Rev. Biochem. 46:473-639.

WILSON, A. C., H. OCHMAN AND E. M. PRAGER. 1987. Molecular time scale for evolution. Trends Genet. 3:241-247.

WING, S. L. AND L. D. BOUCHER. 1998. Ecological aspects of the Cretaceous flowering plant radiation. Annu. Rev. Earth Planet. Sci. 26:379-421.

XU, S. Z. 2000. Phylogenetic analysis under reticulate evolution. Mol. Biol. Evol. 17:897-907.

YANG, Z. H. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396-1401.

YANG, Z. H. 1994. Maximum likelihood phylogenetic estimation from DNA sequence with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306-314.

YANG, Z. H. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13:555-556.

YANG, Z. H., N. GOLDMAN AND A. FRIDAY. 1995. Maximum-likelihood trees from DNA sequences - a peculiar statistical estimation problem. Syst. Biol. 44: 384-399.

YANG, Z. H., R. NIELSEN, N. GOLDMAN AND A. M. K. PEDERSEN. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

YANG, Z. H. AND B. RANNALA. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo method. Mol. Biol. Evol. 14:717-724.

YANG, Z. H. AND W. J. SWANSON. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol. Biol. Evol. 19:49-57.

ZHANG, W. P. 2000. Phylogeny of the grass family (Poaceae) from *rpl16* intron sequence data. Mol. Phylogenet. Evol. 15:135-146.

ZHARKIKH, A. 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39:315-329.

ZHARKIKH, A. AND W.-H. LI. 1995. Estimation of confidence in phylogeny: The complete-and-partial bootstrap technique. Mol. Phylogenet. Evol. 4:44-63.

ZIMMERMAN, W. 1934. Research on phylogeny of species and of single characters. Am. Nat. 68:381-384.

ZIMMERMAN, W. 1943. Die Methoden der Phylogenetik *In* Die Evolution der Organismen (G. Heberer, ed.). Pp. 20-56. G. Fischer, Jena.

ZUCKERKANDL, E. AND L. PAULING. 1962. Molecular disease, evolution, and genic heterogeneity *In* Horizons in biochemistry (M. Kasha and B. Pullman, ed.)^eds.). Pp. 189-225. Academic Press, New York.

ZUCKERKANDL, E. AND L. PAULING. 1965a. Evolutionary divergence and convergence in proteins *In* Evolving genes and proteins (V. Bryson and H. J. Vogel, ed.)^eds.). Pp. 97-166. Academic Press, New York.

ZUCKERKANDL, E. AND L. PAULING. 1965b Molecules as documents of evolutionary history. J. Theor. Biol. 8:357-366.

APPENDICES

Appendix 3.1 - Bootstrap and jackknife support for clades in a 357 plant-taxa phylogeny

	100	1000 reps	100 reps, 5 trees kept							
Clades	NNI	NNI	NNI	SPR	TBR	No swapping	NNI	SPR	TBR	No swapping
Solanaceae	100	100	100	100	100	100	100	100	100	100
Ipomea/Solanaceae	100	100	100	100	100	99	100	100	100	97
Boraginacaeae	100	100	100	100	100	98	100	100	100	99
Solanales	58	60	56	60	68	44	50	59	60	48
Scrophulariaceae	50	55	50	63	54	50	43	65	64	46
Buddleja/Catalpa	53	52	54	58	41	50	44	52	60	43
Lamiaceae	95	96	95	96	96	95	97	99	98	93
Lamiaceae/Thunbergia	53	50	53	53	55	54	52	51	66	41
Buddleja/Catalpa/Lamiaceae/Scrophulariaceae/Thunbergia/Verbena	91	91	92	89	92	90	94	87	88	83
Lamiales	100	100	100	100	100	98	100	100	99	100
Bouvardia/Rubia	100	100	100	100	100	100	100	100	100	100
Bouvardia/Coffea/Rubia	81	82	83	84	83	75	78	84	92	91
Rubiaceae	91	91	90	93	95	89	86	96	93	91
Apocynaceae	100	100	100	100	100	100	100	100	100	100
Apocynaceae/Strychnos	74	72	78	83	81	68	66	81	76	59
Gentianales	100	100	100	100	100	90	100	100	99	94
Gentianales/Lamiales	64	66	62	67	69	37	58	62	61	40
Gentianales/Lamiales/Solanales	100	100	99	100	100	76	98	100	100	86
Aucuba/Garrya	100	100	100	100	100	100	100	100	100	100
Aucuba/Garrya/Eucommia	72	71	65	71	81	73	75	72	68	68
Aucuba/Garrya/Eucommia/Pyrenacantha	60	55	52	57	58	40	50	66	65	52
euasterids I	40	50	42	60	56	19	33	60	53	17
Apium/Pittosporum	n/a	n/a	43	49	60	52	49	57	56	62
Apium/Hedera/Pittosporum	97	99	100	99	100	96	98	98	100	95
Adoxaceae	88	92	92	96	92	93	90	93	94	89
Campanulaceae	100	100	100	100	100	100	100	100	100	100
Campanulaceae/Roussea	57	54	57	58	66	53	56	71	64	47

Cichorium/Menyanthes	56	57	60	60	58	63	54	62	52	49
Corokia/Phelline	67	50	59	66	59	59	64	52	60	64
Cichorium/Corokia/Menyanthes /Phelline	67	66	63	61	68	68	63	75	66	51
Asterales	83	71	70	80	80	64	74	81	82	58
euasterids II	66	62	66	80	76	41	56	81	78	44
Aquilfoliaceae	100	100	100	100	100	100	100	100	100	99
Aquilfoliaceae/Helwingia	100	99	99	100	100	100	99	100	100	100
Aquifoliales	58	55	50	78	76	72	42	70	69	66
Aquifoliales/euasterids II	53	51	53	76	79	31	45	81	73	36
Euasterids	47	50	52	60	64	17	42	60	51	14
Adinandra/Eurya	99	99	100	100	100	99	99	98	99	99
Ternstroemiaceae	67	67	64	86	79	61	73	81	82	65
Primulaceae	100	100	100	100	100	100	100	100	100	99
Clavija/Primulaceae	72	78	80	73	84	63	72	79	80	67
Clavija/Maesa/Primulaceae	99	100	99	100	100	99	100	100	100	100
Argyrodendron(Planchonella)/ Pouteria	52	n/a	86	86	82	80	87	84	81	80
Sapotaceae	85	91	100	100	100	100	100	100	100	100
Lecythidaceae	100	100	100	100	100	98	100	100	100	100
Lecythidaceae/Sapotaceae	97	99	36	44	52	32	30	56	47	29
Ebenaceae	46	50	100	100	100	100	100	100	100	100
Marcgravia/Tetramerista	100	100	51	53	58	58	64	63	62	51
Impatiens/Margravia/ Tetramerista	53	57	100	100	100	100	100	100	100	100
Theaceae	100	100	95	96	94	82	96	99	93	86
Ericales	96	95	92	96	99	71	96	100	98	76
Cornaceae	95	97	94	96	94	90	91	94	96	89
Hydrangeaceae	83	91	99	100	99	99	97	100	100	99
Cornales	100	99	96	100	98	93	94	98	98	94
Acer/Aesculus	68	75	97	99	100	90	99	97	97	91
Cupaniopsis/Koelreuteria	94	96	100	100	100	98	100	100	100	99
Sapindaceae	95	92	100	100	100	100	100	100	100	100
Simaroubaceae	95	98	100	100	100	100	100	100	100	100
Citrus/Poncirus	100	100	100	100	100	100	100	100	100	100
Citrus/Poncirus/Ruta	100	100	70	73	71	67	70	68	71	69
Citrus/Poncirus/Ruta/ Zanthoxylum	100	100	99	100	100	98	98	100	100	97
Rutaceae	52	n/a	97	98	97	83	95	99	99	91
Meliaceae	100	100	87	93	97	84	88	83	89	86

Meliaceae/Simaroubaeae/ Rutaceae	65	73	95	99	100	92	97	98	97	93
Pistacia/Schinus	100	99	60	51	59	54	57	56	55	51
Anacardiaceae	98	98	100	100	100	100	100	100	100	100
Anacardiaceae/Busera	92	87	67	79	77	57	65	68	67	62
Sapindales	99	97	100	100	100	99	100	100	100	100
Adansonia/Ochroma	74	59	72	69	68	61	83	73	72	67
Adansonia/Bombax/Chorisia/ Dombeya/Gossypium/Ochroma /Sterculia/Tilia	100	100	75	65	77	57	62	64	74	60
Grewia/Theobroma	62	66	91	90	87	85	86	81	84	83
Malvaceae	100	100	100	100	100	100	99	100	100	100
Anisoptera/Sarcolaena	62	74	65	68	58	58	61	63	70	59
Cistaceae	43	n/a	100	100	100	100	100	100	100	100
Anisoptera/Cistaceae/ Sarcolaena	67	70	99	100	100	100	100	100	100	100
Anisoptera/Cistaceae/ Muntingia/Sarcolaena	94	89	44	58	54	42	35	50	49	43
Phaleria/Thymelea	100	100	99	99	99	96	97	98	97	95
Thymeleaceae	54	58	100	100	100	100	100	100	100	100
Bixa/Diegodendron	100	100	98	98	100	97	97	100	100	99
Bixaceae	100	100	79	90	89	77	73	77	82	82
Malvales	44	50	97	100	100	82	96	100	100	84
Brassica/Megacarpea	95	98	96	98	97	97	96	96	98	97
Brassica/Megacarpea/Stanleya	100	100	100	100	100	100	100	100	100	100
Brassicaceae	99	97	91	85	90	97	88	92	88	92
Brassicaceae/Reseda	84	78	98	95	95	96	99	99	95	96
Brassicaceae/Floerkea/Reseda	98	97	100	99	99	98	100	100	99	95
Brassicaceae/Carica/Floerka/ Reseda	96	97	82	90	92	70	79	86	90	69
Brassicales	100	100	98	100	99	79	96	100	100	79
Brassicales/Malvales/ Sapindales	94	91	53	75	83	42	49	78	87	37
Clidemia/Metrosideros/ Vochysia	97	97	81	82	82	68	85	75	84	62
Fuschia/Punica	99	99	90	85	87	83	92	88	90	82
Fuschia/Punica/Quisqualis	86	98	81	85	87	63	87	84	81	59
Myrtales	99	99	100	100	100	96	100	100	100	95
Melianthaceae	51	62	100	100	100	100	99	100	99	100
Francoa/Melianthaceae	70	50	100	100	100	100	100	100	100	100
Geraniaceae	47	56	100	100	100	100	100	100	100	100

Francoa/Geraniaceae/ Melianthaceae	n/a	n/a	52	48	53	36	57	54	54	30
Stachyurus/Staphylea	75	85	100	100	100	100	100	100	100	100
Hippocratea/Plagiopteron	87	94	94	98	97	96	93	95	100	93
Hippocratea/Plagiopteron/ Salacia	76	84	66	78	72	52	67	78	71	47
Celastraceae	100	100	73	64	72	84	72	75	67	88
Celastraceae/Stackhousia	27	50	74	72	74	71	68	70	71	76
Celastraceae/Parnassia/ Stackhousia	100	100	100	100	100	99	100	100	100	100
Afrostyrax/Celastraceae/ Parnassia/Stackhousia	100	100	67	61	68	36	63	70	59	41
Averhoa/Rourea	100	100	100	100	100	100	100	100	100	100
Playtheca/Sloanea	55	52	69	64	72	65	65	73	61	58
Eucryphia/Platytheca/Sloanea	100	100	97	100	99	94	96	99	99	89
Oxalidales	94	96	100	100	100	100	100	100	100	98
Carallia/Erythroxylum	68	60	90	90	95	92	93	91	86	88
Mapighiaceae	n/a	50	100	100	100	100	100	100	100	100
Hymenanthera/Rinorea	74	60	100	100	100	100	100	100	100	99
Linaceae	61	77	100	100	100	100	100	100	100	100
Medusagyne/Ochna	100	100	100	100	100	100	100	100	99	100
Passiflora/Turnera	61	60	100	100	100	100	100	100	100	100
Malpighiales	100	100	91	95	96	80	93	99	95	81
Balanites/Guaiacum	69	66	57	50	54	57	53	60	56	63
Zygophyllaceae	98	96	85	84	85	81	89	83	89	77
Krameria/Zygophyllaceae	100	100	99	100	100	100	99	100	100	97
Betula/Casurina	91	80	85	93	86	82	90	93	89	82
Betula/Casurina/Myrica	100	100	58	67	66	51	64	69	62	44
Betula/Casurina/Myrica/ Pterocarya	100	100	95	98	98	90	93	96	95	87
Fagales	100	100	100	100	100	100	99	100	100	98
Coriaria/Corynocarpus	100	100	93	88	93	76	93	89	94	88
Cucurbitaceae	100	100	100	100	100	100	100	100	100	100
Cucurbitaceae/Datisca	91	92	68	59	71	44	70	62	66	46
Cucurbitales	48	54	99	100	100	98	100	100	100	99
Rosaceae	83	86	85	91	94	90	85	93	88	86
Elaeagnus/Rhamnus	100	100	55	58	64	58	59	64	65	61
Humulus/Trema	84	88	98	100	100	94	99	98	99	97
Morus/Urtica	54	56	87	91	86	76	78	83	87	79
Humulus/Morus/Trema/Urtica	93	95	99	100	100	97	97	100	99	96

Eleagnus/Humulus/Morus/ Rhamnus/Trema/Urtica	100	100	91	89	96	87	92	94	92	79
Rosales	92	93	69	86	88	51	70	86	84	53
Fabaceae	100	100	100	100	100	98	100	100	100	97
Polygalaceae	66	66	100	100	100	100	100	100	100	100
Fabales	100	100	94	93	95	66	90	97	90	70
Cucurbitaceae/Fabales/ Fagales/Rosales	48	50	46	62	54	16	39	60	56	16
eurosids I	88	89	31	55	60	9	32	53	51	9
Picramniaceae	66	66	100	100	100	100	100	100	100	100
rosids	99	99	65	68	81	4	58	74	76	18
Altingiaceae	86	85	100	100	100	100	100	100	100	100
Chrysoplenium/Peltoboykinia	100	100	90	91	92	85	91	88	93	88
Saxifragaceae	97	93	97	99	98	92	98	95	96	94
Itea/Pterstemon	77	73	100	100	100	97	100	100	99	98
Itea/Pterstemon/Saxifragaceae	100	100	50	48	52	42	46	45	48	29
Corylopsis/Hamamelis	100	100	97	98	97	95	93	96	95	99
Hamamelidaceae	86	89	91	95	89	72	96	96	87	87
Dudleya/Sedum	42	50	100	100	99	100	100	100	100	100
Dudleya/Kalanchoe/Sedum	37	50	99	100	99	100	100	100	99	99
Crassulaceae	100	100	100	100	100	100	100	100	100	98
Haloragaceae	61	61	100	100	100	100	100	100	100	100
Halagoraceae/Penthorum	100	100	100	100	100	100	100	100	100	100
Crassulaceae/Haloragaceae/ Penthorum	91	88	80	78	76	69	85	78	76	78
Saxifragales	98	97	59	62	62	18	44	57	64	16
Aetoxicon/Berberidopsis	97	100	97	99	100	99	98	97	96	99
Amaranthus/Spinacia	48	50	100	100	100	99	100	100	100	100
Delosperma/Ercilla	99	96	57	52	69	63	67	63	64	65
Phytolaccaceae	92	93	70	85	73	76	75	77	70	64
Bougainvillea/Phytolaccaceae	100	100	90	85	79	84	95	87	85	88
Bougainvilliea/Limeum/ Phytolaccaceae	100	100	61	68	66	72	68	71	68	77
Bougainvilliea/Limeum/ Phytolaccaceae/Rhipsalis	100	100	83	90	91	91	82	95	86	82
Amaranthus/Bougainvillea/ Limeum/Phytolaccaceae/ Rhipsalis/Spinacia	100	100	58	67	72	54	59	71	63	55
Amaranthus/Bougainvillea/ Limeum/Phytolaccaceae/Rhipsalis/Silene/Spinacia	100	100	100	99	100	97	100	98	100	100

Drosera/Nepenthes	74	83	80	83	81	69	88	76	81	72
Polyganaceae	47	50	100	100	99	99	97	98	96	96
Plumbago/Polygonaceae	97	97	100	100	100	100	100	100	100	100
Drosera/Nepenthes/Plumbago/ Polygonaceae	100	100	66	51	46	48	72	51	46	48
Amaranthus/Bougainvillea/ Drosera/Frankenia/Limeum/ Nepenthes/Phytolaccaceae/ Plumbago/Polygonaceae/ Rhipsalis/Silene/Simmondsia/ Spinacia	62	64	43	63	62	34	47	58	56	
Caryophyllales	76	70	97	100	100	93	96	100	100	93
Dillenia/Schumacheria	88	88	94	97	97	97	97	96	96	96
Dilleniaceae	66	70	100	100	100	100	100	100	100	100
Gunnera/Myrothamnus	81	83	84	87	82	73	77	81	85	77
Santalum/Thesium	72	65	71	71	63	58	62	69	69	62
Opilia/Santalaceae	100	99	100	100	99	98	100	100	100	98
Santalales	55	50	93	88	85	75	82	86	91	73
Vitaceae	n/a	n/a	100	100	100	100	100	100	100	100
asterids/caryophyllids/rosids/ Saxifragales/Vitaceae	85	84	47	64	60	19	41	64	62	12
asterids/caryophyllids/Gunnera/ Myrothamnus/rosids/ Saxifragales/Vitaceae	98	98	84	100	99	62	90	100	100	63
Buxaceae	100	100	99	99	100	100	99	100	100	98
Buxaceae/Didymeles	67	67	97	98	100	91	99	99	100	91
Tetracentron/Trochodendron	49	50	100	100	100	100	100	100	100	100
asterids/Buxaceae/ Caryophyllids/Didymeles/ Gunnera/Myrothamnus/rosids/ Saxifragales/Vitaceae Trochodendraceae	100	97	67	78	84	36	70	88	82	39
Lambertia/Roupala	92	95	75	73	84	86	80	79	77	77
Proteaceae	100	100	100	100	100	100	100	100	100	100
Plantanus/Proteaceae	79	80	79	84	81	80	80	71	76	77
Proteales	63	69	55	49	48	40	65	58	60	39
Proteales/Sabia	100	100	36	57	44	42	42	52	38	38
asterids/Buxaceae/ caryophyllids/Didymeles/ Gunnera/Myrothamnus /Proteales/rosids/Sabia/ Saxifragales/	85	86	66	76	63		64	68	83	28

Trochodendraceae/Vitaceae										
Berberidaceae	100	100	100	100	100	100	100	100	100	100
Glaucidium/Hydrastis	n/a	50	48	46	46	49	41	43	42	53
Ranunculaceae	96	91	71	76	69	86	76	83	82	83
Berberidaceae/Ranunculaceae	99	100	63	78	69	72	73	74	77	63
Berberidaceae/Menispermum/ Ranunculaceae	98	100	93	95	92	80	94	93	93	74
Berberidaceae/Decaisnea/ Menispermum/Ranunculaceae	100	100	92	95	92	89	94	91	91	84
Berberidaceae/Decaisnea/ Euptelea/Menispermum/ Ranunculaceae	39	50	91	94	93	80	94	90	89	74
Ranunculales	75	69	91	98	96	85	92	92	99	77
eudicots	84	84	83	100	98	59	87	99	99	63
Androcymbium/Bomerea	100	100	100	100	100	100	100	100	100	100
Melanthiaceae	84	84	78	80	84	82	83	78	80	72
Androcymbium/Bomerea/ Melanthiaceae	52	60	65	67	55	39	47	56	59	38
Nomocharis/Tulipa	37	n/a	84	89	89	85	84	90	90	87
Lloydia/Nomocharis/Tulipa	68	67	99	98	100	93	99	99	100	96
Liliaceae	100	100	96	96	97	81	96	96	95	84
Lapageria/Liliaceae	48	n/a	100	100	100	100	100	100	100	100
Liliales	74	78	75	86	87	57	64	83	88	49
Anthericum/Asparagus/Ipheion	75	71	28	42	51	44	36	40	42	50
Anthericum/Asparagus/Bulbine/ Ipheion	94	94	64	82	72	73	69	77	86	70
Anthericum/Asparagus/Bulbine/ Ipheion/Xeronema	93	95	61	82	78	66	67	79	77	68
Orchidaceae	86	90	100	100	100	100	100	100	100	99
Barbacenia/Sphaeradenia/ Stemona	90	94	80	93	94	84	77	95	89	81
Blandfordia/Rhodohypoxis	94	89	65	60	70	70	60	59	47	68
Tecophilaeaceae	100	100	100	100	100	99	100	100	100	100
Dioscoreales	82	79	100	100	100	99	100	100	99	100
Poaceae	58	62	100	100	100	100	100	100	100	100
Poales	93	89	99	100	98	92	100	100	100	89
Asparagales/Dioscoreales/ Liliales/Orchidaceae/ Pandanales/Poales	97	99	83	97	99	72	86	100	95	66
Tofieldiaceae	92	97	98	99	100	100	100	99	100	100
Tofieldiaceae/Spathiphyllum	100	100	94	90	96	83	86	91	92	92

Asparagales/Dioscolales/ Liliales/Orchidaceae/ Pandanales/Poales/ Tofieldiaceae	79	80	76	83	83	64	81	93	88	59
monocots	34	50	88	93	93	70	82	92	97	55
Annona/Eupomatia	69	66	68	71	74	59	76	71	74	63
Degeneria/Galbulimia	64	62	64	70	80	58	75	66	73	54
Magnoliaceae	52	n/a	82	88	89	84	87	90	92	90
Degeneria/Galbulimia/ Magnoliaceae	n/a	50	52	58	67	43	66	61	59	42
Annona/Degeneria/Eupomatia/ Galbulimia/Magnoliaceae	100	100	65	56	58	39	56	61	58	56
Magnoliales	52	50	95	100	98	86	97	100	99	93
Aristolochia/Lactoris	86	82	64	63	63	48	65	62	62	43
Asarum/Saruma	58	62	100	100	100	100	100	100	100	100
Aristolochiaceae	41	50	58	55	55	35	60	57	60	38
Saururaceae	100	100	97	96	95	92	92	92	98	97
Piperaceae	99	100	100	100	100	100	100	100	100	100
Piperaceae/Saururaceae	100	100	100	100	100	95	100	100	100	93
Piperales	100	50	68	74	84	46	67	78	76	48
Drimys/Tasmannia	n/a	n/a	87	85	87	87	88	92	90	78
Winteraceae	91	50	100	100	100	100	100	100	100	100
Canellaceae	100	100	100	100	100	100	100	100	100	100
Canellaceae/Winteraceae	87	n/a	84	92	93	86	90	90	94	83
Calycanthaceae	82	83	100	100	100	100	100	100	100	100
Lauraceae	85	86	100	100	100	100	100	100	100	100
Monimiaceae	67	73	100	100	100	100	100	100	100	100
Lauraceae/Monimiaceae	69	69	60	68	49	61	60	68	63	61
Calycanthaceae/Lauraceae/ Monimiaceae	87	87	48	57	44	45	59	51	64	50
Laurales	56	53	92	92	100	72	87	95	93	75
Chloranthus/Sarcandra	49	55	100	99	99	97	100	100	99	99
Chloranthaceae	96	95	95	98	95	91	99	93	94	98
Illicium/Schisandra	57	50	100	100	100	100	100	100	100	100
Austrobaileya/Illicium/ Schisandra	100	100	99	100	100	95	99	98	99	93
Nymphaceae	43	50	100	100	100	100	100	100	100	100
Amborella/Austrobaileya/ Canellaceae/Ceratophyllum/ Chloranthaceae/eudicots/ Illicium/Laurales/Magnoliales/	93	91	99	100	100	98	99	100	100	100

monocots/Nymphaceae/ Piperales/Schisandra/ Winteraceae										
Gnetum/Welwitschia	100	100	77	71	70	81	81	76	75	81
Ephedra/Gnetum/Welwitschia	100	100	100	100	100	99	100	100	100	100
Metasequoia/Taxus	66	69	100	100	100	100	100	100	100	99
Metasequoia/Podocarpus/ Taxus	83	88	79	86	86	79	65	78	83	68
Amborella/Austrobaileya/ Canellaceae/Ceratophyllum/ Chloranthaceae/eudicots/ Ephedra/Ginkgo/Gnetum/ Illicium/Laurales/Magnoliales/ Metasequoia/monocots/ Nymphaceae/Piperales/ Podocarpus/Schisandra/Taxus/ Welwitschia/Winteraceae	100	100	100	100	100	100	100	100	100	100

Appendix 4.1 - Taxonomic distribution (subfamily level), characters and methods used in the reviewed publications that represent the source tree for the grass supertrees.

References	Taxonomic distribution	Character used	Method used
Barker, 1997	Poaceae	Plastid DNA	Parsimony
Barker et al., 1995	Arundinoids	Plastid DNA	Parsimony
Barker et al., 1999	Arundinoids	Plastid DNA	Parsimony
Baum, 1982	Pooids	Morphology	Parsimony
Baum et al., 1986	Pooids	Morphology	Parsimony
Buckler and Holtsford, 1996	Panicoids	Nuclear DNA	Parsimony
Catalan et al., 1997	Pooids	Plastid DNA	Parsimony
Clark et al., 1995	Poaceae	Plastid DNA	Parsimony
Cummings et al., 1994	Poaceae	Plastid DNA	Parsimony
Darbyshire and Warwick, 1992	Pooids	Plastid restriction sites	Parsimony
Davis and Soreng, 1993	Poaceae	Plastid restriction sites	Parsimony
Doebley et al., 1990	Poaceae	Plastid DNA	Parsimony
Duvall and Morton, 1996	Poaceae	Plastid DNA	Parsimony
Duvall et al., 1993	Bambusoids	Plastid restriction sites	Parsimony
Esen and Hilu, 1991	Arundinoids	Immunology	Distances
Esen and Hilu, 1993	Panicoids	Immunology	Distances
Frederiksen and Seberg, 1991	Pooids	Morphology	Parsimony
Gaut et al., 1999	Poaceae	Nuclear DNA	Parsimony
Grebenstein et al., 1996	Pooids	Satellite DNA	Parsimony
Grebenstein et al., 1998	Pooids	Nuclear DNA	Parsimony
Hamby and Zimmer, 1987	Poaceae	Nuclear DNA	Parsimony
Hilu and Esen, 1992	Chloridoids	Immunology	Distances
Hilu and Liang, 1997	Poaceae	Plastid DNA	Parsimony
Hilu and Wright, 1982	Poaceae	Morphology	Distances
Hilu et al., 1999	Poaceae	Plastid DNA	Parsimony
Hsiao et al., 1993	Poaceae	Nuclear DNA	Parsimony
Hsiao et al., 1994	Pooids	Nuclear DNA	Parsimony
Hsiao et al., 1995	Pooids	Nuclear DNA	Distances
Hsiao et al., 1998	Arundinoids	Nuclear DNA	Parsimony
Hsiao et al., 1999	Poaceae	Nuclear DNA	Parsimony

Katayama and Ogihara, 1996	Poaceae	Plastid DNA	Parsimony
Kelchner and Clark, 1997	Bambusoids	Plastid DNA	Parsimony
Kellogg, 1989	Pooids	Morphology	Parsimony
Kellog and Watson, 1993	Poaceae	Morphology	Parsimony
Kellogg, 1992	Pooids	Plastid restriction sites	Parsimony
Kellogg, 1998a	Poaceae	Morphology + plastid DNA	Parsimony
Kellogg, 1998b	Pooids	Nuclear DNA	Parsimony
Kellogg and Appels, 1995	Pooids	Nuclear DNA	Parsimony
Liang and Hilu, 1995	Poaceae	Plastid DNA	Parsimony
MacFarlane and Watson, 1982	Pooids	Morphology	Taxonomy
Mason-Gamer and Kellog, 1996	Pooids	Nuclear DNA + plastid restriction sites	Parsimony
Mason-Gamer et al., 1998	Poaceae	Nuclear DNA	ML
Mathews and Sharrock, 1996	Poaceae	Nuclear DNA	Parsimony
Mathews et al., 2000	Poaceae	Nuclear DNA	Parsimony
Monte et al., 1993	Pooids	Nuclear restriction sites	Parsimony
Morton and Clegg, 1993	Poaceae	Plastid DNA	Parsimony
Nadot et al., 1994	Poaceae	Plastid DNA	Parsimony
Nadot et al., 1995	Poaceae	Plastid DNA	Parsimony
Petersen and Seberg, 1997	Pooids	Plastid DNA	Parsimony
Soreng and Davis, 1998	Poaceae	Plastid DNA	Parsimony
Soreng et al., 1990	Pooids	Plastid DNA	Parsimony
Van den Borre and Watson, 1997	Chloridoids	Morphology	Distances
Verboom et al., 1994	Arundinoids	Morphology	Taxonomy
Watanabe et al., 1994	Bambusoids	Plastid DNA	Parsimony
Watson et al., 1985	Poaceae	Morphology	Taxonomy
Zhang, 2000	Poaceae	Plastid DNA	Parsimony

Appendix 6.1 - Species sequenced for rbcL and/or the trnL-F intron and intergenic spacer regions. Subfamilies, tribes and subtribes arranged follow Clayton & Renvoize (1986). Subfamilies according to GPWG (2001) are also provided: Anomochlooideae An, Arundinoideae Ar, Aristoideae Ari, Bambusoideae B, Centothecoideae Ce, Chloridoideae Ch , Danthoniodeae D, Ehrhartoideae E, Pharoideae Ph, Pooideae Po, Panicoideae Pa, Puelioideae Pu.

Note: the information for species sequenced in Royal Botanic Gardens, Kew have not been obtained yet from our collaborators, but will be available in subsequent publications.

TAXON		NUMBER/GENBANK	VOUCHER
ARUNDINOIDEAE			
Arundineae			
Ar	*Amphipogon strictus* R. Br.		
Ar	*Arundo donax* L.		
Ar	*Arundo donax* L.	UK	
Ar	*Arundo donax* L.	131	*Hodkinson 131*
Ar	*Centropodia glauca* (Nees) T. A. Cope		
D	*Cortaderia richardii* (Endl.) Zotov.	G20	
D	*Cortaderia* sp. Stapf	158	*Hodkinson 158*
D	*Danthonia decumbens* (L.) DC.	K23	*Salamin s.n.*
D	*Danthonia spicata* Roem. & Schult.		
Ar	*Elymus patagonicus* Speg.		
Ar	*Elymus trachycaulus* (Link) Hoover or (Link) Gould ex. Shinnners		
Ar	*Elytrophorus globularis* Hack.		
Ar	*Elytrophorus globularis* Hack.	UK	
	Gynerium sagittatum Beauv.		
Ar	*Hakonechloa macra* (Munro) Makino ex Honda	K24	*Salamin s.n.*
Ar	*Monachather paradoxus* Steud.		
D	*Rytidosperma nudiflorum* (not in IPNA)		
Ar	*Molinia caerulea* (L.) Moench.	12076 UK	
Ar	*Molinia littoralis* (not in IPNA)	K21	*Salamin s.n.*
Ar	*Moliniopsis japonica* Hayata (Syn. of *Molinia*)		
Ar	*Phragmites australis* (Cav.) Steud. or Trin. ex. Steud.		
Ar	*Phragmites australis* (Cav.) Steud. or Trin. ex. Steud.	11081 UK	
Ar	*Phragmites* sp. (Cav.) Steud. or Trin. ex. Steud.	203	*Hodkinson 203*
Ar	*Plinthanthesis paradoxa* (R. Br.) S. T. Blake		
Ar	*Styppeiochloa gynoglossa* (Goossens) de Winter		
Thysanolaneae			
Ce	*Thysanolaena maxima* (Roxb.) Kuntze		

Ce	*Thysanolaena maxima* (Roxb.) Kuntze	K10	*Salamin s.n.*

Aristideae

Ari	*Aristida congesta* subsp. *barbicollis* Roem. & Schult.		
Ari	*Aristida congesta* subsp. *barbicollis* Roem. & Schult.	UK	
Ari	*Stipagrostis zeyheri* subsp. *zeyheri* (Nees) de Winter		

BAMUSOIDEAE

Bambuseae

Arundinariinae

B	*Ampelocalamus scandens* Hsueh & W.D.I. Li	18B	*Kew 1991-1157*
B	*Arundinaria oedogonata* (not in IPNA)	1988	
B	*Arundinaria tecta* Muhl.	1992	
B	*Arundinaria alpina* K. Schumann	80B	*S.Philips s.n.*
B	*Arundinaria gigantea* (Walter) Muhlenberg	94B	*MWC 1995*
B	*Aulonemia longiaristata* Clark & Londono	78B	*L.Clark, P.Asimb.1389.*
B	*Bashania gingchengshanensis* Keng & Yi	162B	
B	*Bashania fargesii* (Camus) Keng & Yi	181B	
B	*Chimonobambusa quadrangularis* (Fenzi) Makino	6B	*Kew 1988-3398*
B	*Chimonobambusa marmorea* (Mitford) Makino	3B	*Kew 1973-20180*
B	*Chimonobambusa marmorea* (Mitford) Makino	1982	
B	*Chusquea circinata* T. R. Soderstrom & C. E. Calderon		
B	*Chusquea culeou* E. Desv.	1975	
B	*Chusquea coronalis* Soderstrom & Calderon	73B	*Stapleton 1126*
B	*Chusquea delicatula* Hitchc.	19B	*Kew 1985-8243*
B	*Drepanostachyum falcatum* (Nees) Keng	12B	*Kew 1996-1426*
B	*Fargesia murieliae* (Gamble) Yi	61B	*Kew 1973-20162*
B	*Fargesia dracocephala* Yi	63B	*Kew 1989-1914*
Pu	*Guaduella marantifolia* (not in IPNA)	UK	
Pu	*Guaduella marantifolia* (not in IPNA)		
B	*Himalayacalamus cupreus* Stapleton14B		*Stapleton s.n.*
B	*Himalayacalamus falconeri* (Hooker ex Munro) Keng	106B	*Mike Bell s.n.*
B	*Himalayacalamus hookerianus* (Munro) Stapleton	17B	*Kew 1973-12236*
B	*Indocalamus tesselatus* var. *hamadeae* (Hatusima) Rifat ex Ohrnb.	29B	*Kew 1991-1532*
B	*Indocalamus latifolius* (Keng) McClure	1994	
B	*Indocalamus tessellatus* (Munro) Keng	33B	*Kew 1973-14424*
B	*Neurolepis elata* (Kunth) Pilger	83B	*L.Clark et al. 1409*
B	*Oligostachyum oedogonatum* (Wang & Ye) Zheng & Huang	139B	*Kew 1994-1243*
B	*Olmeca* sp. Soderstrom	1986	
B	*Olmeca recta* Soderstrom	160B	*Kew 1983-2575*
B	*Otatea acuminata* (Munro) C.E. Calderon & Soderstrom	34B	*Kew 1992-3550*

B	*Pleioblastus viridistriatus* (Regal) Makino	59B	*Kew 1984-3078*
B	*Pleioblastus linearis* (Hackel) Nakai	97B	*MWC1998*
B	*Pleioblastus pygmeus var. distichus* (Miquel) Nakai	60B	*Kew 1982-1222*
B	*Pseudosasa amabilis* (McClure) P.C. Keng	62B	*Kew 1995-3859*
B	*Pseudosasa japonica* (Siebold & Zuccarini ex Steudel) Makino ex Nakai	151B	*Kew 1973-14386*
B	*Pseudosasa japonica* (Siebold & Zuccarini ex Steudel) Makino ex Nakai	1985	
B	*Qiongzhuea macrophylla* Wen & Ohrnb.	56B	*Stapelton 1124*
B	*Qiongzhuea tumidissinoda* Hsueh & Yi ex Ohrnb.	67B	*Kew 1989-315*
B	*Sasa palmata* E. G. Camus	1991	
B	*Sasa palmata* f. *nebulosa* (Makino) Suzuki	11B	*Kew 1973-14422*
B	*Sasa ramosa* (Makino) Makino	64B	*Kew 1973-80178*
B	*Sinobambusa tootsik* (Siebold ex Makino) Makino	1996	
B	*Sinobambusa tootsik* (Siebold ex Makino) Makino	7B	*Kew 731-1243*
B	*Thamnocalamus spathiflorus* Munro	1978	
B	*Thamnocalamus spathiflorus var. crassinodus* (Yi) Stapleton	138B	*Kew 1960-64402*
B	*Thamnocalamus tessellatus* (Nees) Soderstrom & Ellis		*Kew 1984-2277*
B	*Yushania anceps* (Mitf.) Lin	13T	*Kew 1990-1414*
B	*Yushania maling* (Gamble) R.B. Majumdar	24B	*Kew 1982-8559*

Melocaninae

B	*Melocanna baccifera* (Roxb.) Kurtz ex Skeels	1739	
B	*Melocanna baccifera* (Roxb.) Kurtz ex Skeels	95B	*MWC1739*
B	*Schizostachyum funghomii* McClure	1408	
B	*Schizostachyum funghnomii* McClure	102B	*MWC1408*
B	*Schizostachyum caudatum* Backer ex Heyne	164B	*Kew 1980-2111*
B	*Schizostachyum zollingerii* Steudel	100B	*MWC1983*
B	*Pseudostachyum polymorphum* Munro	98B	*MWC1934*

Bambusinae

B	*Arthrostylidium* sp. Rupr.	53B	*Stapleton 1133*
B	*Bambusa glaucescens* (Wild.) E. D. Merill	1420	
B	*Bambusa valida* (Q. H. Dai) D. Ohrnberger	1743	
B	*Bambusa emeiensis* L. C. Chia & H. L. Fung	1393	
B	*Bambusa multiplex 'Alphonose Karr'* (Loureiro) Raeuschel ex Schultes & Schultes	137B	
B	*Bambusa multiplex* var.*gracilli*ma (Loureiro) Raeuschel ex Schultes & Schultes	30B	*Kew 1990-1427*
B	*Bambusa vulgaris* Shrader ex Wendland	135B	*Kew 1973-21090*
B	*Brachystachyum densiflorum* (Rendle) Keng	54B	*Stapleton 1134*
B	*Dendrocalamus barbatus* C. J. Hsueh & D. Z. Li	1562	
B	*Dendrocalamus brandisii* (not in IPNI)	1563	
B	*Dendrocalamus giganteus* Munro	165B	*Stapleton 452*
B	*Dendrocalamus membranaceus* Munro	147B	*Kew 1992-3549*
B	*Gigantochloa verticillata* (Steudel) Widjaja	140B	*Kew 1973-12238*

B	*Hibanobambusa tranquillans* (Kuidzumi) Maruyama & Okamura	149B	*Hodkinson s.n.*
B	*Hibanobambusa tranquillans ‘Shiroshima’* (Koidzumi) Maruyama & Okamura	150B	*Hodkinson s.n.*
B	*Neomicrocalamus andropogonifolius* (Griff) Stapleton	9B	*Kew 1991-3178*
B	*Oreobambos buchwaldii* Schmann	105B	*Kare s.n.*
B	*Phyllostachys bambusoides* Sieb. ex. Zucc.		
B	*Phyllostachys dulcis* McClure	1965	
B	*Phyllostachys flexousa* (Carriere) Riviere & Riviere	35B	*Kew 1973-14404*
B	*Phyllostachys nigra var henoni*s (Mitford) Muroi	134B	*Kew 1973-20509*
Pu	*Puelia ciliata* Franch.		
Pu	*Puelia ciliata* Franch.	UK	
B	*Racemobambos hepburnii* Dransfield	41B or 88B	67K
B	*Rhipidocladum harmonicum* (Parodi) McClure	82B	L.*Clark et al. 1103*
B	*Semiarundinaria fastuosa* (Marliae ex Mitford) Makino ex Nakai	22B	*Kew 1973-20176*
B	*Semiarundinaria yamadorii* Muroi	5B	*Kew 1985-2082*
B	*Shibataea kumasaca* (Steudel) Makino ex Nakai	4B	*Kew 1984-3084*
B	*Shibataea chinensis* Nakai	2B	*Kew 1994-1217*
B	*Sinocalamus oldhamii* (Munro) McClure	55B	*Stapleton 1135*
B	*Thyrsostachys siamensis* Gamble	1551	
B	*Thyrostachys siamensis* Gamble	96B	*MWC1411*

Others

B	*Borinda perlonga* Stapleton	146B	*Kew 1995-4215*
B	*Borinda emeryi* Stapleton	8B	*Kew 1992-0401A*
B	*Gaoligongshania megalothyrsa* (Handel-Mazzetti) Li, Hseuh & Xia	50B	

Ehrharteae

E	*Ehrharta erecta* Lam.	G18	*G. Hodkinson 18*
E	*Ehrharta erecta* Lam.	G25	*G. Hodkinson 25*

Olyreae

B	*Lithachne humilis* Soderstrom		
B	*Olyra latifolia* L.	77B	*A.M. deCervalho 4394*
B	*Raddia brasiliensis* Bertol.	74B	*Kew s.n.*
B	*Raddia portoi* Kuhlm.	79B	*A.M. deCervalho 4360*

Oryzeae

E	*Leersia oryzoides* (L.) Sw. or Michx.		
E	*Oryza sativa* L.		
E	*Oryza sativa* L.	UK	
E	*Oryza sativa* L.	46	*Hodkinson 46*
E	*Oryza sativa* L.	*Genbank*	*Genbank*

E	*Zizania texana* Hitchc.		

Parianeae

B	*Pariana parvispica* R. Pohl	528	*Hodkinson 528*

Phaenospermateae

Po	*Phaenosperma globosa* Munro ex. Benth.	G11	

Phareae

Ph	*Pharus latifolius* L.	514	*Hodkinson 514*
Ph	*Pharus* sp. L.	578	*Hodkinson 578*

Streptocheteae

An	*Streptochaeta sodiroana* Hack.	574	*Hodkinson 574*
An	*Streptochaeta sodiroana* Hack.	575	*Hodkinson 575*

Anomochloeae

An	*Anomochloa marantoidea* Brongn.		

CENTOTHECOIDEAE

Centotheceae

Ce	*Centotheca lappacea* Desv.	235	
Ce	*Chasmanthium latifolium* (Michx.) Yates		
Ce	*Chasmanthium latifolium* (Michx.) Yates	K15	*Salamin s.n.*

CHLORIDOIDEAE

Cynodonteae

Boutelouinae

Ch	*Bouteloua gracilis* Steud.	G17	

Chloridinae

Ch	*Chloris argentina* Lillo & Parodi	K20	*Salamin s.n.*
Ch	*Cynodon dactylon* (L.) Pers.	K4	*Salamin s.n.*
Ch	*Cynodon transvalensis* Burrt Daty	116	*Hodkinson 116*
Ch	*Spartina pectinata* Bosc ex. Link	G6	
Ch	*Spartina pectinata* Bosc ex. Link		
Ch	*Spartina pectinata* Bosc ex. Link		
Ch	*Spartina anglica* C.E. Hubbard	153	*Hodkinson 153*

Zoysiinae

Ch	*Perotis* sp. Ait	274	

Ch	*Tragus racemosus* (L.) All.	K31	*Salamin s.n.*
Ch	*Zoysia japonica* Steud.	K13	*Salamin s.n.*

Eragrostideae

Sporobolinae

Ch	*Eragrostis capensis* Trin. or Jedwabnick		
Ch	*Calamolvilfa* sp. Hack.	132	
Ch	*Calamolvilfa longifolia* Hack.	K27	*Salamin s.n.*
Ch	*Crypsis schoenoides* (L.) Lam.	K6	*Salamin s.n.*
Ch	*Sporobolus indicus* (L.) R.Br	K19	*Salamin s.n.*
Ch	*Sporobolus fertilis* (Steud) W.D. Clayton	K22	*Salamin s.n.*
Ch	*Muhlenbergia mexicana* Schreb (not in IPNI)	119	
Ch	*Muhlenbergia lindheineri* Hitchc.	K32	*Salamin s.n.*
Ch	*Muhlenbergia racemosa* Britton, Sterns & Poggenb	K1	*Salamin s.n.*
Ch	*Muhlenbergia racemosa* Britton, Sterns & Poggenb.	G29	
Ch	*Muhlenbergia* sp. Schreb	K3	*Salamin s.n.*

Eleusininae

Ch	*Eleusine tristachya* Kunth	K26	*Salamin s.n.*
Ch	*Eleusine indica* Steud.	126	*Hodkinson 126*
Ch	*Eleusine coracana* (L.) Gaertn.	127	*Hodkinson 127*
Ch	*Eragrostis virescens* J & C. Presl.	K12	*Salamin s.n.*
Ch	*Eragrostis curvula* Nees	K25	*Salamin s.n.*

Leptureae

Ch	*Lepturus repens* R.Br.	272	

Poppophoreae

Ch	*Enneapogon scaber*		
Ch	*Enneapogon polyphyllus* (Domin) N.T. Burb.	K9	*Salamin s.n.*

Other chloridoid check position

D	*Merxmuellera macowanii* (Stapf) Conert		
D	*Merxmuellera macowanii* (Stapf) Conert	K29	*Salamin s.n.*

PANICOIDEAE

Paniceae

Cenchrinae

Pa	*Cenchrus setigerus* Steud. or Vahl.		
Pa	*Cenchrus incertus* M. A. Curtis	K28	*Salamin s.n.*
Pa	*Cenchrus incertus* M. A. Curtis	123	*Hodkinson 123*

Digitariinae

Pa	*Digitaria sanguinalis* (L.) Scop.	110	*Hodkinson 110*

Setariinae

Pa	*Echinochloa crus-galli* (L.) Beauv (*frumentacea*)	125	*Hodkinson 125*
Pa	*Panicum* sp. L.	565	*Hodkinson 565*
Pa	*Panicum virgatum* L.	120	*Hodkinson 120*
Pa	*Panicum virgatum* L.	G34	
Pa	*Paspalum dilatatum* Steud.	128	*Hodkinson 128*
Pa	*Paspalum notatum* Fluegge	K8	*Salamin s.n.*
Pa	*Paspalum quadrifarium* Lam.	K18	*Salamin s.n.*
Pa	*Pennisetum alopecuroides* Steud.	K17	*Salamin s.n.*
Pa	*Pennisetum glaucum* (L.) R. Br.		
Pa	*Pennisetum macrourum* Trin.	117	*Hodkinson 117*
Pa	*Setaria italica* (L.) P. Beauv.		

Neurachninae

Pa	*Neurachne tenuifolia* S. T. Blake		

Arundinelleae

Pa	*Tristachya biseriata* Stapf. or Chiov		

Andropogoneae

Andropogoninae

Pa	*Andropogon gerardii* Vit.	G2	
Pa	*Andropogon gerardii* Vit.	75	*Hodkinson 15*
Pa	*Arthraxon* Beauv.	G3	
Pa	*Arthraxon* Beauv.	111	*Hodkinson/Nicolas*
Pa	*Cymbopogon citratus* Stapf.	129	*Hodkinson 129*
Pa	*Schizachyrium scoparium*	113	*Hodkinson/Salamin*

Saccharinae

Pa	*Spodiopogon sibiricus* Trin. *(or Eriochrysis 114)*	*128*	*Lancaster 210*
Pa	*Eulalia irritans* (R. Br.) Kuntze	137	*Adams 1756*
Pa	*Eulalia villosa* (Thunb.) Nees	132	*Devenish 1282*
Pa	*Eulalia quadrinervis* (Hack.) Kuntze	134	*Polunin et al. 3294*
Pa	*Eulalia tripsicata* (Schlut.) Henrard	138	*Clarkson 10062*
Pa	*Imperata cylindrica* P.Beauv. Raeuschel	G14	
Pa	*Imperata cylindrica* P.Beauv. Raeuschel	122	*Marsden 3*
Pa	*Miscanthus floridulus* (Labill.) Warb. ex K. Schum. & Lauterb.		
Pa	*Miscanthus giganteus* Greef & Deuter ex Hodkinson & Renvoize		
Pa	*Miscanthus violaceus*		
Pa	*Miscanthus sorghum*		
Pa	*Miscanthus nepalensis* (Trin.) Hack.	25	*Hodkinson 1*
Pa	*Miscanthus oligostachyus* Stapf.	16	*Hodkinon 13*
Pa	*Miscanthus oligostachyus* Stapf.	161	*Hodkinson 161*
Pa	*Miscanthus oligostachyus* Stapf.		
Pa	*Miscanthus sacchariflorus* (Maxim) Benth. & Hook	5791	*Renvoize 5791*

Pa	*Miscanthus sacchariflorus* (Maxim) Benth. & Hook		
Pa	*Miscanthus sacchariflorus* (Maxim) Benth. & Hook	7343	
Pa	*Miscanthus sacchariflorus* ‘ Purpurascens’ (Maxim) Benth. & Hook	61	*Hodkinson s.n. 1987-272*
Pa	*Miscanthus sinensis* Anderss.	G9	
Pa	*Miscanthus sinensis* Anderss. subsp. *condensatus*	7	*Renvoize s.n. 1969-19091*
Pa	*Miscanthus sinensis* Anderss.	5	*Hodkinson 40*
Pa	*Miscanthus sinesnis* Anderss. ' Yakushimanum'	63	*Hodkinson 21*
Pa	*Miscanthus sinensis* Anderss.	30	*ADAS MB94/07*
Pa	*Miscanthus transmorrisonensis* Hayata	65	*Hodkinson 20*
Pa	*Miscanthidium teret*		
Pa	*Miscanthidium junceus* (Stapf.) Pilger		
Pa	*Saccharum contortum* L.		
Pa	*Saccharum officinarum* L. (sugarcane cv.)	104	*Kew 1973-12242*
Pa	*Saccharum ravennae* Beauv.		
Pa	*Saccharum spontaneum* L.		
Pa	*Sclerostachya fuscus* Roxb.		
Pa	*Spodiopogon sibiricus* Trin.	G25	
Pa	*Spodiopogon sibricus* Trin.		

Ischaeminae

Pa	*Karroochloa purpurea* (L. f.) Conert & Turpe		

Sorghinae

Pa	*Sorghum halepense* (L.) Pers.	G38	
Pa	*Sorghum halepense* (L.) Pers.	6	*Hodkinson 10*
Pa	*Sorghum caf.*	130	*Hodkinson 130*

Anthristiriinae

Pa	*Hyparrhenia hirta* Stapf.		
Pa	*Themeda triandra* Forsk.	MWC9286	*Salamin s.n.*

Tripsacinae

Pa	*Tripsacum dactyloides* L. or Schlecht	G31	
Pa	*Zea diploperennis* Iltis, Doebley & Guzman	164	*Hodkinson 164*
Pa	*Zea mays* L.		

POIDEAE

Aveneae

Alopecurinae

Po	*Agrostis canina* L.	G1	
Po	*Agrostis stolonifera* L.	10744 UK	
Po	*Alopecurus geniculatus*	10745 UK	
Po	*Alopecurus pratensis* L.	G30	
Po	*Alopecurus pratensis* L.	30	*Hodkinson 30*
Po	*Ammophila breviligulata* Fernald	G27	
Po	*Lagarus ovatus* L.	6	*Hodkinson 6*

Po	*Calamagrostis epigejos* Huds. or Kar & Kir or Roth. or Steud.	G4	
Po	*Phleum pratense* L.	10801 UK	
Po	*Phleum pratense* L.	G24	

Aveninae

Po	*Aira elegantissima* Schur.	G37	
Po	*Aira praecox* L.	3	*Hodkinson 3*
Po	*Avena fatua* L.	31	*Hodkinson 31*
Po	*Avena sativa* L.		
Po	*Arrhenatherum elatius* (L.) Beauv. ex. J. & C. Presl	10747 UK	
Po	*Arrhenatherum elatius* (L.) Beauv. ex. J. & C. Presl	G10	
Po	*Arrhenatherum elatius* (L.) Beauv. ex. J. & C. Presl.	27	*Hodkinson 27*
Po	*Deschampsia* sp. P. Beauv.	32	*Hodkinson 32*
Po	*Deschampsia caespitosa* (L.) Beauv.	5	*Hodkinson 5*
Po	*Helictotrichon requienii* (Mutel) Henrard	G35	
Po	*Helictotrichon* Schult. sp.		
Po	*Holcus lanatus* L.	25	*Hodkinson 25*
Po	*Koeleria pyramidata* Beauv.	G13	
Po	*Koeleria cristata* (L.) Pers.10b	10	*Hodkinson 10*
Po	*Koeleria* sp. (L.) Pers.		
Po	*Trisetum flavescens* (L.) Beauv.	10879 UK	
Po	*Trisetum flavescens* (L.) Beauv.	18	*Hodkinson 18*
Po	*Trisetum paniceum* Pers.	G8	

Phalaridinae

Po	*Anthoxanthum odoratum* L.	10824 UK	
Po	*Hierochloe odorata* Beauv. or Britton, Stern & Pogg	G16	
Po	*Anthoxanthum odoratum* L.	2	*Hodkinson 2*
Po	*Phalaris arundinacea* L.	G15	
Po	*Phalaris arundinacea* L.	20	*Hodkinson 20*

Bromeae

Po	*Bromopsis erecta* Fourr. (*Bromus* Synon.)	10832 UK	
Po	*Bromus inermis* Stev.		
Po	*Bromus commutatus* Bieb. or Schrad. or Guss ex. Steud.	10751 UK	
Po	*Bromus ramosus* Huds.	41	*Hodkinson 41*
Po	*Bromus* sp. L.	1	*Hodkinson 1*

Lygeeae

Po	*Lygeum spartum* L.	18	*Hodkinson 18*

Meliceae

Po	*Glyceria fluitans* R. Br.	10776 UK	
Po	*Glyceria maxima* (Hartm.) Holmb.	19	*Hodkinson 19*
Po	*Melica macra* Nees.	5578	

Po	*Melica uniflora* Retz.	10857 UK	
Po	*Melica uniflora* Retz.	44	*Hodkinson 44*
Po	*Melica* sp.	G12	

Nardeae

Po	*Nardus stricta* L.	11075 UK	

Poeae

Po	*Ampelodesmos mauritanica* (not in IPNA)	5523	
Po	*Briza media* L.	10831 UK	
Po	*Briza media* L.	12	*Hodkinson 12*
Po	*Catapodium rigidum* (L.) C.E. Hubbard	23	*Hodkinson 23*
Po	*Cynosurus cristatus* L.		
Po	*Dactylis marina* Burrill	G22	
Po	*Dactylis glomerata* L.	26	*Hodkinson26*
Po	*Festuca rubra* agg	10769 UK	
Po	*Festuca rubra* subsp *juncea*	G7	
Po	*Festuca rubra*	15	
Po	*Festuca rubra* L.		
Po	*Lamarckia aurea* (L.) Moench.	G26	
Po	*Lolium perenne perenne* L.	10790 UK	
Po	*Lolium* L.	29	*Hodkinson 29*
Po	*Puccinellia distans* (L.) Parl.		
Po	*Poa trivialis*	10867 UK	
Po	*Poa trivialis* L.	28	*Hodkinson 28*
Po	*Sesleria caerulea*	G32	
Po	*Sesleria caerulea* (L.) Ard.	45	*Hodkinson 45*
Po	*Vulpia ciliata* (Pers.) Link or St. Lager	G33	

Stipeae

Po	*Nassella trichotoma* (Nees) Hack. ex. Arechav. or Hack. ex Arechav	G40	
Po	*Oryzopsis* Michaux	2	
Po	*Oryzopsis* Michaux	66	
Po	*Stipa 79* L.	79	
Po	*Stipa579* L.	579	*Hodkinson 579*
Po	*Stipa dregeana var dregeana* Steud.		
Po	*Stipa gigantantea* Lag or Ledeb. or Link	5576	

Triticeae

Po	*Elymus patagonicus* Speg.		
Po	*Elymus trachycaulus* (Link) Hoover or (Link) Gould ex. Shinnners		
Po	*Brachypodium sylvaticum* (Huds.) Beauv.	22	*Hodkinson 22*
Po	*Hordeum brachyantherum* Nevski or R. Phil or R. Regel		
Po	*Hordeum jubatum* L.		

Po	*Hordeum lechleri* (Steud.) Schenck		
Po	*Hordeum secalinum* Gus or Schreb or Savi	10779	
Po	*Leymus chinensis* (Trin.) Tsvelev		
Po	*Leymus arenarius* (L.) Hochst.	7	*Hodkinson 7*
Po	*Pseudoroegneria spicata* (Pursh) A. Löve (Elymus syn.)		
Po	*Triticum aestivum* L.		
Po	*Triticum aestivum* L.	AF148757	*Briggs et al., 2000*
Po	*Triticum baeoticum*	AF519168	*Briggs et al., 2000*
Po	*Thinopyrum scirpeum*	AF519167	*Briggs et al., 2000*
Po	*Lophopyrum elongatum*	AF519166	*Briggs et al., 2000*
Po	*Pseudoroegneria spicat*	AF519160	*Briggs et al., 2000*
Po	*Elymus canadensis*	AF519131	*Briggs et al., 2000*
Po	Elymus virginicus	AF519144	*Briggs et al., 2000*
Po	*Heteranthelium pilifer*	AF519153	*Briggs et al., 2000*
Po	*Taeniatherum caput med*	AF519164	*Briggs et al., 2000*
Po	*Aegilops speltoides*	AF519112	*Briggs et al., 2000*
Po	*Eremopyrum orientale*	AF519151	*Briggs et al., 2000*
Po	*Agropyron mongolicum*	AF519117	*Briggs et al., 2000*
Po	*Australopyrum retrofra*	AF519118	*Briggs et al., 2000*
Po	Henrardia persica	AF519152	*Briggs et al., 2000*
Po	*Peridictyon sanctum*	AF519154	*Briggs et al., 2000*
Po	Psathyrostachys juncea	AF519170	*Briggs et al., 2000*
Po	*Secale cereale*	AF519162	*Briggs et al., 2000*
Po	*Hordeum murinum*	AF519126	*Briggs et al., 2000*
Po	*Hordeum jubatum*	AF519123	*Briggs et al., 2000*

OUTGROUPS

ECDEIOCOLACEAE

Ecdeiocolea monostachya F. .Muell.	AF148734	*Briggs et al., 2000*
Georgeantha hexandra B.G. Briggs & L. A. S. Johnson	AF148733	*Briggs et al., 2000*

FLAGERELEACEAE

Flagellaria indica L.	AF206769	*Soltis, Soltis, Chase*

JOINVILLEACEAE

Joinvillea ascendens Gaudich.		
Joinvillea plicata (Hook. f.) Newell & Stone	L01471	*Duvall et al. unpub.*

RESTIONACEAE

Baloskion gracile (R. Br.) B. G. Briggs & L. A. S. Johnson	AF148764	*Briggs et al., 2000*
Elegia cuspidata Mart.	AF148774	*Briggs et al., 2000*

OTHERS

Spartochloa scirpoidea (Steud.) C. E. Hubbard	*Mike Fay*
Bothriochloa caucasia (Trin.) C. E. Hubbard	112
Cyperochloa hirsuta M. Lazarides & L. Watson	
Cyperochloa hirsuta M. Lazarides & L. Watson	UK

Appendix 6.2 - ML estimates of parameters for A) the HKY85+Γ *model of substitution for* rbcL *and B) the GTR+I+*Γ *model of substitution for* trnLF *and combined data sets.*

A)

Parameters	Values
A	0.271
C	0.192
G	0.248
T	0.289
ti/tv ratio	1.718
Gamma shape	0.221

B)

Parameters		Data sets	
		trnLF	combined
A		0.336	0.297
C		0.167	0.184
G		0.166	0.218
T		0.331	0.301
R matrix	AC	0.785	0.945
	AG	1.487	2.051
	AT	0.751	0.916
	CG	0.833	0.827
	CT	1.596	2.753
	GT	1.000	1.000
invariant sites		0.008	0.416
Gamma shape		0.388	0.833