Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

A duplicate gene rooting of seed plants and the phylogenetic position of flowering plants

Sarah Mathews, Mark D. Clements and Mark A. Beilstein

Phil. Trans. R. Soc. B 2010 **365**, 383-395
doi: 10.1098/rstb.2009.0233

Supplementary data

"Data Supplement"
http://rstb.royalsocietypublishing.org/content/suppl/2010/01/13/365.1539.383.DC1.html
"Audio Supplement"
http://rstb.royalsocietypublishing.org/content/suppl/2010/01/19/365.1539.383.DC2.html

References

This article cites 99 articles, 50 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/365/1539/383.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/365/1539/383.full.html#related-urls

Subject collections

Articles on similar topics can be found in the following collections

evolution (488 articles)
taxonomy and systematics (45 articles)

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

To subscribe to *Phil. Trans. R. Soc. B* go to: **http://rstb.royalsocietypublishing.org/subscriptions**

Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2010) **365**, 383–395
doi:10.1098/rstb.2009.0233

A duplicate gene rooting of seed plants and the phylogenetic position of flowering plants

Sarah Mathews*, Mark D. Clements and Mark A. Beilstein

Arnold Arboretum of Harvard University, 22 Divinity Avenue, Cambridge, MA 02138, USA

Flowering plants represent the most significant branch in the tree of land plants, with respect to the number of extant species, their impact on the shaping of modern ecosystems and their economic importance. However, unlike so many persistent phylogenetic problems that have yielded to insights from DNA sequence data, the mystery surrounding the origin of angiosperms has deepened with the advent and advance of molecular systematics. Strong statistical support for competing hypotheses and recent novel trees from molecular data suggest that the accuracy of current molecular trees requires further testing. Analyses of phytochrome amino acids using a duplicate gene-rooting approach yield trees that unite cycads and angiosperms in a clade that is sister to a clade in which *Gingko* and Cupressophyta are successive sister taxa to gnetophytes plus Pinaceae. Application of a cycads + angiosperms backbone constraint in analyses of a morphological dataset yields better resolved trees than do analyses in which extant gymnosperms are forced to be monophyletic. The results have implications both for our assessment of uncertainty in trees from sequence data and for our use of molecular constraints as a way to integrate insights from morphological and molecular evidence.

Keywords: seed plant rooting; phytochromes; amino acids; duplicate genes

1. INTRODUCTION

> As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever-branching and beautiful ramifications.
>
> (Charles Darwin, *The origin of species*, 1859)

Flowering plants represent the most significant branch in the tree of land plants, with respect to the number of extant species, their impact on the shaping of modern ecosystems (Friis *et al.* 1987) and their economic importance. Nonetheless, an understanding of their origin remains elusive (Crepet 2000; Stockey *et al.* 2009). Unlike so many persistent phylogenetic problems that have yielded to insights from DNA sequence data, the mystery surrounding the origin of angiosperms has deepened with the advent and advance of molecular systematics. Results from these studies have suggested that angiosperms have no near relatives among living gymnosperms (cycads, *Gingko*, conifers and gnetophytes), contradicting inferences from morphology that place angiosperms near gnetophytes (Parenti 1980; Crane 1985; Doyle & Donoghue 1986; Loconte & Stevenson 1990; Nixon *et al.* 1994; Doyle 2006, 2008; Hilton & Bateman 2006). Nonetheless, strong statistical support for competing hypotheses and recent novel trees from molecular data (e.g. Chumley *et al.* 2008; Rai *et al.* 2008) suggest that the accuracy of current molecular trees requires further testing. Without a consensus regarding the relationships of angiosperms with other seed plants, the task of determining where flowers came from remains formidable. In addition to the need for accurate estimates from sequence data, morphological evidence must play a significant role since so many seed plant groups, including candidate sister groups of angiosperms, are extinct. Thus, it is particularly important to understand how insights from molecular data have influenced our interpretation of the morphological evidence, and to grapple with the problem of how to integrate insights from the two types of data.

In the analyses described in this paper, we explore a nuclear gene dataset to determine whether the species tree implied by the inferred tree is consistent with published hypotheses of seed plant phylogeny inferred from DNA sequence data. Specifically, we analyse amino acid data from three phytochrome loci, *PHYA*/*N*, *PHYB*/*P* and *PHYC*/*O*, to test the monophyly of living gymnosperms and the position of gnetophytes. The results of these analyses support the position of gnetophytes as sister to Pinaceae, consistent with results from many other sequence analyses; however, they do not support the monophyly of extant gymnosperms and in this respect they contradict most molecular trees (reviewed in Mathews 2009). Instead, the data suggest a novel rooting of the seed plant tree such that two clades are resolved, one that includes *Ginkgo*, conifers and gnetophytes and one that includes cycads and angiosperms. To explore the implications of this alternative rooting on our assessment of morphological

* Author for correspondence (smathews@oeb.harvard.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2009.0233 or via http://rstb.royalsocietypublishing.org.

One contribution of 16 to a Discussion Meeting Issue 'Darwin and the evolution of flowers'.

This journal is © 2010 The Royal Society

Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

evidence, we analyse a published morphological dataset (Doyle 2008), employing the molecular result as a backbone constraint.

(a) *Issues in the inference of seed plant phylogeny*

Issues and limitations in the inference of seed plant phylogeny from nucleotide data have been discussed in four recent papers (Burleigh & Mathews 2004, 2007*a*,*b*; Mathews 2009) and will be touched on just briefly here. Divergences among the living lineages of seed plants occurred from 350 to 150 Myr ago (Magallón & Sanderson 2005). One of the limitations of DNA sequence data for the inference of ancient divergences is the small number of character states (four) at single nucleotide positions, because the greater are the divergences among terminal taxa, the more likely it is that multiple substitutions per site will have occurred along the branches leading to those taxa, potentially leading to substitutional saturation and a loss of information about changes. Thus, to reconstruct accurately the changes at individual sites, it is critical to sample taxa that break up long branches. This is a particular problem in the inference of seed plant phylogeny because taxa that could be sampled for this purpose are extinct (for recent analyses that highlight the diversity of extinct taxa, see Doyle (2006, 2008) and Hilton & Bateman (2006)). Not only does this increase the difficulty of accurate reconstruction of relationships among ingroup taxa, it also increases the likelihood that the branch from the outgroup will attach to one of the long ingroup branches, perhaps with high statistical support, regardless of whether this position is correct. The latter phenomenon may account for a rooting of the seed plant tree that places the gnetophytes as sister to all other seed plants (Sanderson *et al.* 2000; Burleigh & Mathews 2004). In these cases, adding more characters, but not more taxa, may exacerbate the error (Felsenstein 1978).

(b) *Analyses of amino acids to infer seed plant phylogeny*

Analyses of amino acids rather than nucleotides may help reduce error introduced by substitutional saturation. For protein-coding genes, the rate of nucleotide substitution is greater than the rate of amino acid substitution owing to the redundancy of the genetic code; on a given tree, evolution at the amino acid level will be more conservative overall than evolution at the nucleotide level, which may be useful for resolving deep nodes. Another potential advantage of protein alignments is the greater number of possible character states (20), which may decrease the error associated with multiple, undetected substitutions per site (e.g. Steel & Penny 2000). However, this potential advantage may be offset in maximum-likelihood (ML) analyses, where the use of amino acid data may greatly increase the number of parameters to be estimated, leading to higher sampling variances. To reduce the number of parameters to be estimated a fixed amino acid transition rate matrix may be used; however, while it is clear that the choice of the amino acid transition model impacts the results of a likelihood analysis (e.g. Whelan & Goldman 2001; Le & Gascuel 2008), the available models are not necessarily the most appropriate for a specific dataset. Bayesian analysis of amino acid sequences is appropriate in this case because it does not rely on a fixed model (Huelsenbeck *et al.* 2008). Estimation of an amino acid rate matrix that is specific to the protein to be analysed is also appropriate. In the analyses reported here, we use a transition matrix inferred from a phytochrome alignment of 97 full- or nearly full-length sequences from across land plants. We are not here advocating the general use of amino acids over nucleotides. Rather, the opportunity to explore the utility of amino acid sequences for a particular problem is an advantage since it is known that multiple nucleotide substitutions result in the failure to detect some proportion of changes on a tree, that rapidly and slowly evolving nucleotide sites in seed plant datasets have conflicting phylogenetic signal (Chaw *et al.* 2000; Frohlich & Parker 2000; Magallón & Sanderson 2002; Rydin *et al.* 2002; Soltis *et al.* 2002; Burleigh & Mathews 2004; Hajibabaei *et al.* 2006), and that the use of amino acid models may improve phylogenetic accuracy by better accounting for dependencies among coding sequence sites (Whelan 2008).

(c) *Phytochromes for a duplicate gene-rooting approach*

In most angiosperms, phytochrome genes occur as a small family comprising three to five members. The completely characterized family of *Arabidopsis* has five genes, *PHYA* through *PHYE* (Sharrock & Quail 1989; Clack *et al.* 1994). Three of these (*PHYA–C*) are members of monophyletic gene lineages that were established before the origin of flowering plants, and they occur in nearly all flowering plants (there is one known absence of *PHYC*, from the published genome of *Populus trichocarpa*). Extant gymnosperms also have three monophyletic phytochrome lineages, and these were established before their origin. It is clear that gymnosperm *PHYP* is an orthologue of angiosperm *PHYB*, while gymnosperm *PHYN* and *PHYO* are putative orthologues of *PHYA* and *PHYC*, respectively (Schmidt & Schneider-Poetsch 2002; Mathews 2006). Most analyses suggest that angiosperm *PHYA* and gymnosperm *PHYN* are orthologous, but they fail to support the orthology of angiosperm *PHYC* and gymnosperm *PHYO* and the position of the *PHYC/O* clade as sister to the *PHYA/N* clade. Instead, *PHYC* and *PHYO* are often resolved as successive sister groups to *PHYA* + *PHYN* (Mathews 2006). A gene tree topology in which *PHYC* and *PHYO* are orthologues, shown in figure 1, requires no hypotheses of undetected gene duplications and losses, and it suggests that the duplications leading to the three lineages *PHYP/B/E*, *PHYN/A* and *PHYO/C* were established early in the history of seed plants, prior to the radiation of crown seed plants. This topology also supports the monophyly of extant gymnosperms, as has been inferred in many analyses of DNA sequence data (Bowe *et al.* 2000; Chaw *et al.* 2000; Nickrent *et al.* 2000; Gugerli *et al.* 2001; Magallón & Sanderson 2002; Rydin & Källersjö 2002; Burleigh & Mathews 2004, 2007*a*,*b*; de la Torre *et al.* 2006; Hajibabaei *et al.* 2006; Wu *et al.* 2007; McCoy *et al.* 2008).

Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

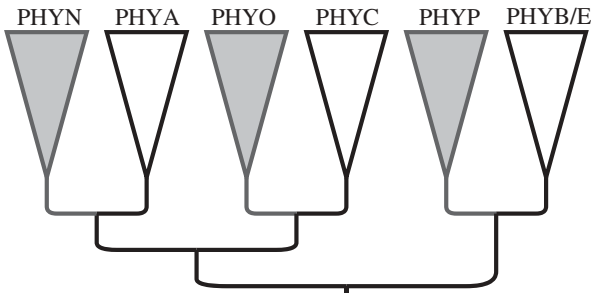

Figure 1. Putative phytochrome gene phylogeny that supports the orthology of gymnosperm PHYN, PHYO and PHYP (grey shading) with angiosperm PHYA, PHYC and PHYB, respectively, and that supports the monophyly of extant gymnosperms.

2. MATERIAL AND METHODS

(a) *Inference of the phytochrome amino acid transition matrix*

Ninety-three full-length or nearly full-length phytochrome sequences from seed plants (electronic supplementary material, appendix 2) and four from free-sporing species (*Selaginella* X61458, *Adiantum PHY1* AB016151 and *PHY2* AB016232, and *Psilotum* X74930) were extracted from a working phytochrome alignment containing 598 sequences. When translated into amino acid data, the 97-sequence matrix, including alignment gaps, contained 1059 aligned sites. Using methods similar to Le & Gascuel (2008), Whelan & Goldman (2001) and Abascal *et al.* (2007), we estimated a 20×20 phytochrome-specific exchangeability matrix, Q, using ML (Felsenstein 1981) and an interative process. (i) We first assumed a fixed phylogenetic tree for phytochromes that was estimated by the maximum parsimony (MP) analysis of the nucleotide data. (ii) We optimized branch lengths and a gamma-shape parameter, with four discrete rate categories (Yang 1993), on this tree, assuming empirical amino acid frequencies and the LG (Le & Gascuel 2008) model of amino acid substitution. (iii) The LG-optimized gamma and branch lengths were then used to estimate a new Q matrix, Q1, for phytochrome sequences. (iv) We then performed another round of parameter optimization (branch lengths and gamma shape) using the Q1 matrix. (v) We continued with parameter optimization and Q estimation until there were no improvements in parameter values, Q exchangeability coefficients and the ML score. Convergence of this iterative procedure occurred after two rounds of Q estimation (Whelan & Goldman 2001; Le & Gascuel 2008). All analyses were performed using R statistical software (R Development Core Team 2009). Data matrices and phylogenetic trees were entered and manipulated in R using the ape package (Paradis *et al.* 2004) and ML analyses in R used the phangorn package (Schliep 2009).

(b) *Phytochrome phylogenetic matrices*

From the same 598-sequence phytochrome alignment, a matrix was prepared for phylogenetic analyses by deletion of sequences from Zygnematales, *Marchantia*, mosses, *Selaginella*, ferns and *Equisetum*, and by omitting in some cases multiple species from genera within angiosperms and gymnosperms. Sequences from eudicots were also deleted from the alignment, except for those from Ranunculales, Proteales, Sabiaceae, Buxales and Trochodendrales. Similarly, sequences from several monocot lineages were deleted, except for those from Acorales and Alismatales. The nucleotide data were translated into amino acids and the resulting alignment comprised 1118 amino acid sites and 380 phytochrome amino acid sequences (PHY), 230 from angiosperms (70 PHYA, 74 PHYC, 69 PHYB and 17 PHYE, a duplicate of PHYB that arose early in extant angiosperms) and 150 from gymnosperms (58 PHYN, 72 PHYP, and 20 PHYO). A total of 166 species were sampled in this matrix (electronic supplementary material, appendix 1). There is a great deal of length variation among the sequences; 9.2 per cent are full length, while most of the rest are 40–60% complete, with a handful of fragments being only 30 per cent complete (electronic supplementary material, appendix 1). The majority of full-length sequences were either downloaded from GenBank or were obtained by using 5′ and 3′ rapid amplification of cDNA ends (RACE; Frohman *et al.* 1988) or thermal asymmetric interlaced PCR (TAIL PCR; Liu & Chen 2007). Since the GenBank accessions are exclusively from eudicots or grasses (angiosperms), and represent gymnosperms very poorly (just four Pinaceae full-length PHY sequences are in GenBank), RACE and TAIL PCR were used to obtain full- or nearly full-length sequences from the angiosperm genera *Nymphaea*, *Ceratophyllum*, *Austrobaileya*, *Schisandra*, *Sarcandra*, *Aristolochia*, *Piper*, *Calycanthus*, *Liriodendron*, *Euptelea*, *Meliosma*, *Acorus* and *Sagittaria*, and from the gymnosperm genera *Zamia*, *Ginkgo*, *Ephedra*, *Gnetum*, *Cephalotaxus*, *Sciadopitys* and *Sequoidendron*, focusing on PHYA, PHYC, PHYN and PHYO. Full-length PHYA, PHYB and PHYE sequences from *Aquilegia* were kindly provided by Elena Kramer (Harvard University) prior to their publication. To test the results obtained from the 380-sequence matrix, a matrix of 119 more complete sequences was also analysed. Just 43 of the 119 sequences were also in the 380-sequence matrix, reflecting the different taxonomic sampling in the two matrices. The 119-sequence matrix has poor coverage of gymnosperms, and of clades that diverge early from the rest of angiosperms; conversely, it has greater coverage of taxa that are well nested in the eudicot and monocot clades.

(c) *Morphological data*

The morphological dataset of Doyle (2008) was kindly provided by Jim Doyle. After the deletion of *Archaefructus* and related characters, the morphological matrix comprised 35 fossil and living taxa and 133 characters. Character definitions and taxon scorings are discussed in Doyle (2006, 2008).

(d) *Phylogenetic analyses*

ML analyses of the phytochrome data were conducted using RAxML v. 7.0.4 (Stamatakis 2006) on the Odyssey cluster at Harvard University. Heuristic and bootstrap searches employed the PHY-specific amino

Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

acid rate matrix, and heuristic searches of the 380-sequence matrix were run multiple times to check for convergence in likelihood values. Bootstrap searches were done separately from the search for the best tree, and employed the 'thorough' rather than the 'fast' bootstrap option, specified by using the '-f i' switch in the configuration file rather than the '-f a' switch, as described in the RAxML documentation. MP analyses of the morphological data were conducted using PAUP* (Swofford 2002). Heuristic and bootstrap analyses (1000 replicates) used 10 random taxon addition replicates, tree bisection and reconnection branch swapping, holding five trees, and saving multiple MP trees. Constrained MP analyses employed the same search settings and enforced a backbone constraint tree of living taxa that united cycads and angiosperms; all other nodes in the constraint tree were unresolved. Bayesian analyses of the morphological data were conducted in MrBayes 3.1 (Huelsenbeck & Ronquist 2001; Ronquist & Huelsenbeck 2003) using the standard discrete model for transitions between character states. The Metropolis-coupled Markov chain Monte Carlo consisted of two independent runs of five million generations; one tree in every 1000 trees was sampled. Output was evaluated using Tracer (Rambaut & Drummond 2007).

(e) *Topology tests*

Approximately unbiased (AU) tests (Shimodaira 2002, 2008) were conducted using the R (http://www.r-project.org/) package, scaleboot, to test whether the phytochrome data could reject an alternative gene topology, or could reject species topologies that have been supported in other seed plant analyses. The tested topologies are: (i) the PHY topology ((PHYA, PHYC)(PHYN, PHYO)), which would imply separate phytochrome duplications, leading to PHYN + PHYO in gymnosperms and PHYA + PHYC in angiosperms, rather than a single duplication leading to PHYN/A and PHYO/C, as shown in figure 1; (ii) monophyly of extant gymnosperms; (iii) (*Ginkgo*(cycads(angiosperms))); (iv) ((cycads, *Ginkgo*)angiosperms); (v) (*Ginkgo*, angiosperms); (vi) (gnetophytes, angiosperms); (vii) (gnetophytes, cupressophytes); (viii) (gnetophytes, conifers), and (ix) (gnetophytes, all other seed plants).

3. RESULTS

(a) *Phytochrome amino acid phylogeny*

ML analysis of the amino acid data yields an optimal tree with three gene lineages that coalesce near the origin of extant seed plants (figure 2 and electronic supplementary material, figure S1). One lineage includes all gymnosperm PHYP and angiosperm PHYB and PHYE sequences (100% bootstrap value). The tree is rooted on the branch to this clade, consistent with all analyses that include seedless plants (Mathews 2006). This lineage is sister to the other two, one of which includes all gymnosperm PHYN and angiosperm PHYA sequences (100% bootstrap value) and one that includes all gymnosperm PHYO and angiosperm PHYC (84% bootstrap value). This suggests that gymnosperm PHYO and

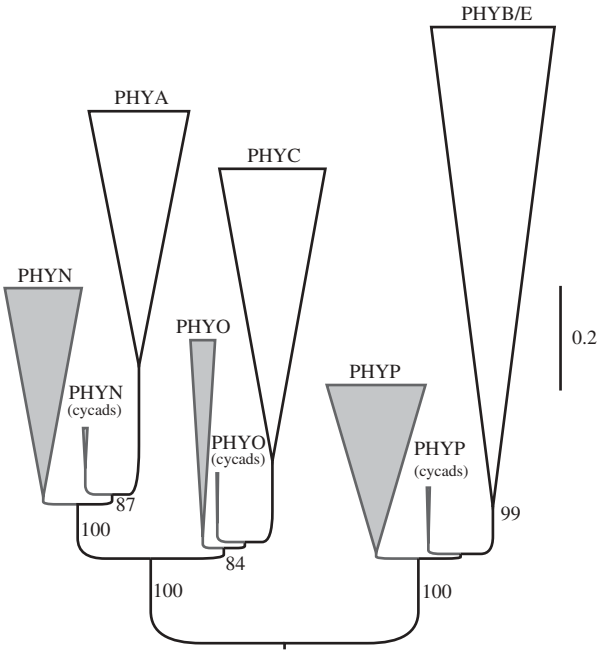

Figure 2. Optimal gene phylogeny for angiosperm and gymnosperm phytochromes that was inferred in ML analyses of the 380-sequence phytochrome amino acid matrix using the PHY-specific amino acid rate matrix. The $-\ln L = 84272.061$. Branch support is shown for selected nodes. The monophyly of PHYA and PHYC is supported by bootstrap values of 100%. A detailed tree is available in the electronic supplementary material, figure S1.

angiosperm PHYC are orthologues, as are gymnosperm PHYN and angiosperm PHYA, and that three major PHY lineages were established before the origin of extant seed plants. The species topologies within the gene clades are congruent (figure 2): cycads and angiosperms are sister taxa, and they are sister to the remaining extant gymnosperms. Within the latter clade, gnetophytes and Pinaceae are sister taxa in the PHYN clade (98% bootstrap value) and the PHYP clade (65% bootstrap value) (data not shown and electronic supplementary material, figure S2). PHYO has not been detected in gnetophytes (Mathews 2006). The topological congruence of the gene clades allows the gene tree to be folded into a species phylogeny without invoking undetected gene duplications and losses (figure 3). The sister-group relationship of cycads and angiosperms is well supported by the PHYN/A data (87% bootstrap value), but not by the PHYO/C or the PHYP/B/E data (less than 50% bootstrap value). Nonetheless, together the results suggest that the monophyly of extant gymnosperms is less certain than has been implied by recent analyses of sequence data, and that the tree of extant seed plants might be rooted between *Gingko* and cycads. Given the large amount of missing data in the 380-sequence matrix, we wished to determine whether analyses of a more complete matrix would yield similar results with respect to the species phylogeny. In the bootstrap consensus from the analysis of 119 mostly full-length sequences (electronic supplementary material, figure S2), *Zamia* PHYN is sister to PHYA (71% bootstrap value) while *Zamia* PHYO is in a polytomy with PHYC and a clade of

Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

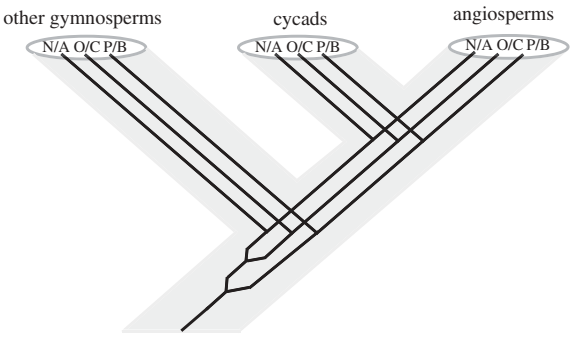

Figure 3. The optimal PHY tree in figure 2 can be folded into a species tree without invoking undetected duplications or losses.

the remaining PHYO (data not shown). *Zamia* PHYP is in a clade with gymnosperms, nested within a clade of Pinaceae and gnetophytes, but support for gymnosperm monophyly is low (58% bootstrap value). A clade of *Gnetum* and *Ephedra* PHYP is in a polytomy with *Zamia* and Pinaceae PHYP, while *Gnetum* and *Pinus* PHYN are sister sequences (98% bootstrap value) (data not shown).

In AU tests (Shimodaira 2002, 2008), three topologies were rejected by both datasets: the PHY topology implying separate duplications in gymnosperms and angiosperms, a sister-group relationship between gnetophytes and angiosperms, and a sister-group relationship between gnetophytes and all other seed plants (table 1). Gymnosperm monophyly could not be rejected by either the 119- or the 380-sequence dataset (table 1). Some topologies could be rejected by the 119-sequence dataset, but not by the 380-sequence dataset (table 1). This may result from the high proportion of missing data in the latter, since the test is based on the estimation of site-wise likelihoods. Assuming that the tests based on the 119-sequence dataset have the most power, the results indicate that the PHY data reject all species topologies tested except those that suggest a rooting on the branch to angiosperms or between cycads and *Ginkgo*, and those that suggest a sister-group relationship between gnetophytes and Pinaceae (table 1).

To explore the implications of a rooting between cycads and *Ginkgo* for the interpretation of morphological evidence, we used it as a backbone constraint in analyses of the morphological dataset constructed by Doyle (2008). Unconstrained MP analyses of the data retrieved the same 40 trees of 346 steps that were obtained by Doyle (2008). The strict consensus of all 40 trees resolves only four clades within crown seed plants: angiosperms, gnetophytes, Podocarpaceae + Pinaceae, and Taxaceae + Cupressaceae (fig. 2 in Doyle 2008). However, the trees fall into two islands and the strict consensus trees for these are reasonably well resolved, particularly with respect to relationships among the living and extinct taxa (fig. 3 in Doyle 2008). Analysis of the data with a backbone constraint in which cycads and angiosperms are sister taxa allows unconstrained taxa to attach to the backbone in an optimal position with respect to the morphological data. In contrast to the fully resolved molecular constraint used by Doyle (2008), in which extant

Table 1. Log likelihoods of optimal tree and trees constrained to the indicated topology, and results from AU topology tests. (*p*-corrected *p*-values by Akaike weights averaging; values greater than 5.0% (bold font) indicate topologies that are not significantly different from the best tree. In gymnoP monophyly trees, PHYP is monophyletic, while cycad PHYN and PHYO are sister to angiosperm PHYA and PHYC, respectively).

topology	likelihood	*p*
119-sequence matrix		
best (gymnoP monophyly)	−69031.31770	**78.54**
cycads + angiosperms	−69042.21268	**27.66**
gymno monophyly	−69035.69901	**43.48**
((PHYA, PHYC) (PHYN, PHYO))	−69140.08546	0
anthophyte	−69099.92565	0.08
ginkgo + angiosperms	−69064.33610	3.88
((cycad + ginkgo)(angios))	−69074.25946	0.54
ginkgo(cycads(angios))	−69064.54910	4.21
(gnetophytes, cupressophytes)	−69066.68142	1.91
(gnetophytes, conifers)	−69070.55370	1.39
(gnetophytes, all other seed plants)	−69093.18006	0.55
380-sequence matrix		
best (cycads + angiosperms)	−84272.06104	**75.97**
gymnoP monophyly	−84273.51451	**59.78**
gymno monophyly	−84280.75512	**47.01**
((PHYA, PHYC) (PHYN, PHYO))	−84385.35086	0
anthophyte	−84442.35106	0.01
ginkgo + angiosperms	−84313.28715	**5.82**
((cycad + ginkgo)(angios))	−84311.36169	**10.99**
ginkgo(cycads(angios))	−84301.40346	**14.25**
(gnetophytes, cupressophytes)	−84302.85687	**8.74**
(gnetophytes, conifers)	−84304.72176	**5.90**
(gnetophytes, all other seed plants)	−84363.21190	0.04

gymnosperms were monophyletic and gnetophytes were sister to Pinaceae, we used a minimal constraint, enforcing just a clade of cycads and angiosperms. MP analysis of the data with this constraint yielded 20 trees of 346 steps. The strict consensus of these trees (figure 4) is identical to the consensus of the trees in one of the two islands found in the unconstrained search. After the divergence of medullosans from the rest of the crown seed plants, there is a major dichotomy leading, on one hand, to a clade of gnetophytes and conifers that is subtended by the fossil conifer *Emporia*, then Ginkgoales, then Cordaitales; this larger clade is in a polytomy with corystosperms, *Autunia* and *Peltaspermum*. On the other hand, there is a clade of angiosperms subtended by *Caytonia*, then Bennettitales, glossopterids + *Pentoxylon*, then Cycadales. Thus, the application of this simpler constraint resulted in a greater resolution, perhaps resulting from reduced conflict between the morphological data and the molecular topology, and it provided a criterion for choosing between trees in two most parsimonious islands. Parsimony bootstrap values on the nodes in the tree from the constrained analysis are low (figure 4). However, an unconstrained Bayesian analysis of the data provides support for

Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

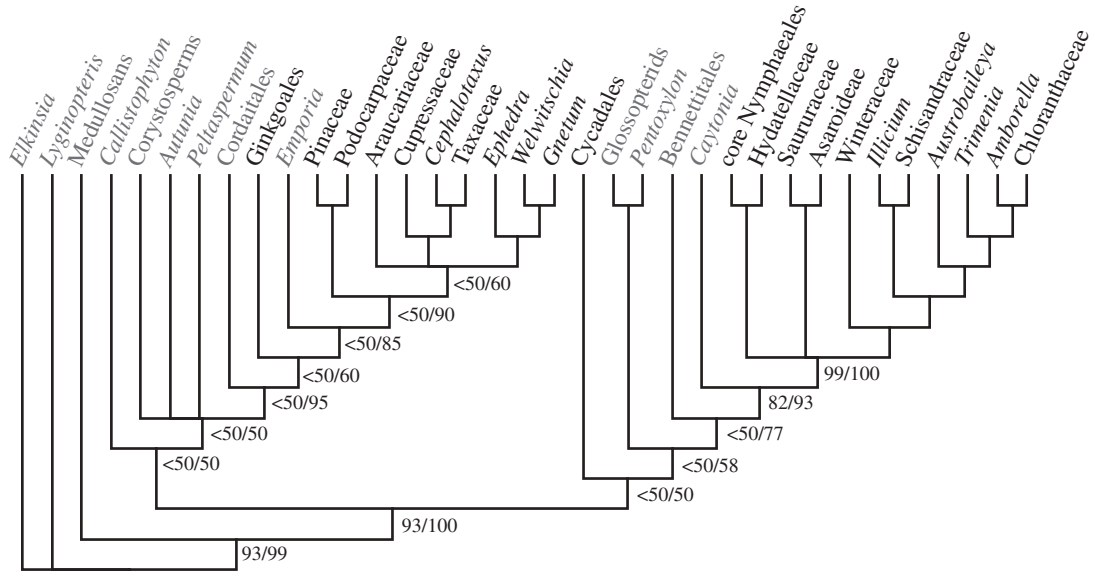

Figure 4. Strict consensus of 20 trees of 346 steps that were inferred in the MP analyses of the Doyle (2008) morphological dataset. The optimal PHY topology was enforced as a backbone constraint. Bootstrap values from constrained analysis (leftmost number) and posterior probabilities (rightmost number) from unconstrained analysis are on the backbone nodes.

certain elements in the strict consensus of trees from the constrained analysis (figure 4). A posterior probability of 0.95 supports a coniferophyte clade (*sensu* Chamberlain 1935), comprising Cordaitales, Ginkgoales, living and extinct conifers and gnetophytes, while a posterior probability of 0.93 supports the position of *Caytonia* as sister to the angiosperms. Gnetophytes are in a clade with conifers in the Bayesian consensus tree (posterior probability of 0.90).

4. DISCUSSION

(a) *Rooting of extant seed plants*

In their suggestion of an alternative rooting of the tree of living seed plants, the phytochrome amino acid analyses challenge an apparently well-supported consensus favouring the monophyly of living gymnosperms. The monophyly of both angiosperms and living gymnosperms has been supported at very high levels in all recent trees from DNA sequences except for those that are rooted on the gnetophyte branch; in the trees in which both angiosperms and gymnosperms are monophyletic, cycads and *Ginkgo* are successive sister taxa to a clade of conifers and gnetophytes (Burleigh & Mathews 2004, 2007*a*; de la Torre *et al.* 2006; Hajibabaei *et al.* 2006; Wu *et al.* 2007; McCoy *et al.* 2008). Donoghue & Doyle (2000) suggested that a rooting of molecular trees on the branch between cycads and *Ginkgo* might be particularly difficult to infer because this branch is very short relative to the branches on which the outgroup branch usually attaches, that is, on the branch to angiosperms or to gnetophytes (Burleigh & Mathews 2004). Tests of the ability of the long angiosperm or gnetophyte branches to attract (e.g. Huelsenbeck *et al.* 1996, 1998) have not been conducted, nor have there been extensive experiments to test alternative rootings with parametric bootstrapping, as there have been with angiosperms (Zanis *et al.* 2002), or to test the rooting signal when different outgroups are used (e.g. Graham & Iles 2009). Nonetheless, a rooting along the branch to gnetophytes is viewed as erroneous since likelihood analyses, which are expected to better detect multiple substitutions on a branch, rarely retrieve this result, nor do parsimony analyses when rapidly evolving sites are excluded (Chaw *et al.* 2000; Magallón & Sanderson 2002; Rydin *et al.* 2002; Soltis *et al.* 2002; Burleigh & Mathews 2004, 2007*a*; Hajibabaei *et al.* 2006; but see Burleigh & Mathews 2007*a*; Rai *et al.* 2008). Despite the lack of empirical data suggesting that branch-length effects are operating, there is value in (i) exploring alternative approaches for inferring the root and (ii) determining whether the signal in a nuclear dataset corroborates trees that have been inferred from mostly organellar data.

(b) *Topology of the phytochrome tree*

The duplicate gene-rooting approach has been used to root the angiosperm tree (Mathews & Donoghue 1999), where outgroups are particularly divergent, as well as the entire tree of life (e.g. Gogarten *et al.* 1989; Iwabe *et al.* 1989; Brown & Doolittle 1995; Baldauf *et al.* 1996; Lake *et al.* 2009), where outgroups are unknown. In the case of the angiosperms, it proved to be a particularly efficient approach, leading to a result that was corroborated by traditional analyses of larger, multi-locus datasets (Parkinson *et al.* 1999; Qiu *et al.* 1999; Barkman *et al.* 2000; Graham & Olmstead 2000; Soltis *et al.* 2000; Zanis *et al.* 2002). Simultaneous analysis of duplicate genes is expected to yield an unrooted network of gene clades; if the species tree topology is the same in each clade, the network can be folded such that it fits into a rooted species tree, without invoking hypotheses of undetected gene duplication or loss (Slowinski *et al.* 1997; Donoghue & Mathews 1998). Inspection of the optimal tree from the analysis of the 380-sequence matrix shows that the PHY clades can be folded

Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

onto one another into a single rooted species tree in which cycads are sister to angiosperms and together they are sister to the other extant gymnosperms (figure 3). There is clear support in the PHYN/A data for a rooting between cycads and *Ginkgo*; in a variety of permutations of the large dataset, which alter taxon and gene sampling, a clade of cycad PHYN sequences is uniformly supported as sister to a clade of angiosperm PHYA sequences. Moreover, PHYO/C and PHYP/B/E clades in the optimal tree are congruent with the PHYA/N topology. These results suggest a rooting not seen in other seed plant phylogenies inferred from sequence data, although it has been suggested by certain morphological analyses (Doyle & Donoghue 1987; Doyle 2006, 2008). The PHY amino acid data also yield a parsimonious solution for the gene phylogeny that has not emerged from analyses of nucleotide data (Mathews 2006), suggesting that the amino acid data are more appropriate for reconstructing deep divergences. Still, since the living gymnosperms are monophyletic in so many other molecular trees, and since the PHY data cannot reject gymnosperm monophyly, one could argue that it is favoured by the weight of evidence. Nonetheless, it remains to be determined whether any of the other datasets can reject the optimal PHY topology for seed plants.

(c) *The phytochrome tree and morphological evidence*

In evaluating the credibility of the rooting suggested by the PHY topology, it is equally important to consider whether or not the rooting between cycads and *Ginkgo* is more or less consistent with morphological evidence than are trees in which gymnosperms are monophyletic. Recent morphological analyses suggest that cycads are sister to a clade that includes angiosperms, *Caytonia*, Bennettitales, glossopterids and *Pentoxylon* while *Ginkgo* is sister to a clade of conifers and gnetophytes (Doyle 2008), or that cycads and *Ginkgo* are successive sister taxa to a clade that includes all other living seed plants (Crane 1985; Doyle & Donoghue 1986; Loconte & Stevenson 1990; Doyle 1996, 2006, 2008; Hilton & Bateman 2006). These trees cannot be trimmed in such a way that gymnosperms are monophyletic, and constraining morphological analyses to make them so can result in a much reduced resolution and/or in a consensus tree that breaks the constraints (e.g., Hilton & Bateman 2006; Doyle 2008). In contrast, one island of trees from a recent morphological analysis is congruent with the species tree rooting implied by the optimal PHY tree (fig. 2 in Doyle 2008). Not surprisingly then, using the PHY topology to constrain analyses of this same morphological dataset favours this island of trees, thus providing a criterion for choosing among most parsimonious trees. Conversely, congruence of the PHY topology with a set of morphological trees increases the credibility of the species topology suggested by the PHY data. Notably, a total evidence tree from morphological and plastid photosystem gene data supports the same rooting (S. Magallón submitted).

(d) *The position of gnetophytes*

In contrast to the novel rooting result, the position of gnetophytes in the phytochrome amino acid trees (electronic supplementary material, figures S1 and S2) is consistent with molecular trees that are rooted on the branch to the angiosperms, which place them in a clade with conifers. Gnetophytes are nested in conifers in these trees, as sister to Pinaceae (Burleigh & Mathews 2004, 2007*a*; de la Torre *et al.* 2006; Hajibabaei *et al.* 2006; Wu *et al.* 2007; McCoy *et al.* 2008) or as sister to cupressophytes (Nickrent *et al.* 2000; Chumley *et al.* 2008); more rarely, gnetophytes are sister to all conifers in so-called 'gnetifer' trees (Hamby & Zimmer 1992; Chaw *et al.* 1997; Stefanović *et al.* 1998; Rydin & Källersjö 2002; Soltis *et al.* 2002; Burleigh & Mathews 2007*a*). Many analyses cannot distinguish among these topologies, because of limited sampling within conifers, but do unite gnetophytes with whichever taxon represents conifers in a particular analysis, usually *Pinus* (e.g. de la Torre *et al.* 2006; Wu *et al.* 2007; McCoy *et al.* 2008). An unusual, and well-supported, topology to emerge in a recent analysis (Rai *et al.* 2008) is depicted in figure 5*a*, as an unrooted tree, along with the seven possible rooted trees. The optimal ML tree from this analysis of 17 plastid genes and associated non-coding regions is rooted on the gnetophyte branch (figure 5*b*), which may be unlikely (Burleigh & Mathews 2004, 2007*a*,*b*). The topology in figure 5*e* is congruent with the species tree suggested by the PHY amino acid data, and it is interesting to note that the rooted tree that is consistent with the monophyly of extant gymnosperms (e.g. figure 5*c*) would support the rarely seen gnetifer topology, as would four of the other possible rootings of the tree (figure 5*d*–*g*). However, this topology is refuted in analyses of a much larger plastid dataset, one that was assembled from 83 plastid genes, which yield trees uniting gnetophytes with cupressophytes (Chumley *et al.* 2008). The PHY data reject topologies that place gnetophytes as sister to cupressophytes or to all conifers (table 1), and this discrepancy between results from plastid and nuclear data bears testing with additional nuclear data. They also reject placing gnetophytes with angiosperms in an anthophyte clade or as sister to all other seed plants. Thus, despite disagreement regarding exactly how gnetophytes are related to conifers in molecular trees, the sequence data have been unanimous in supporting a link between the two groups, and this is consistent with certain characters that are shared by gnetophytes and conifers (Bailey 1944; Bierhorst 1971; Carlquist 1996; Doyle 1996). Nevertheless, gnetophytes and Bennettitales, members of the anthophyte clade in many morphological trees (Crane 1985; Doyle & Donoghue 1986; Nixon *et al.* 1994; Rothwell & Serbet 1994; Doyle 2006, 2008; Hilton & Bateman 2006; Friis *et al.* 2007), share potential seed, pollen and anatomical synapomorphies (Friis *et al.* 2007, 2009).

(e) *Insights for future analyses*

The analyses presented here represent an exercise with a set of single genes. What can we learn about species

Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

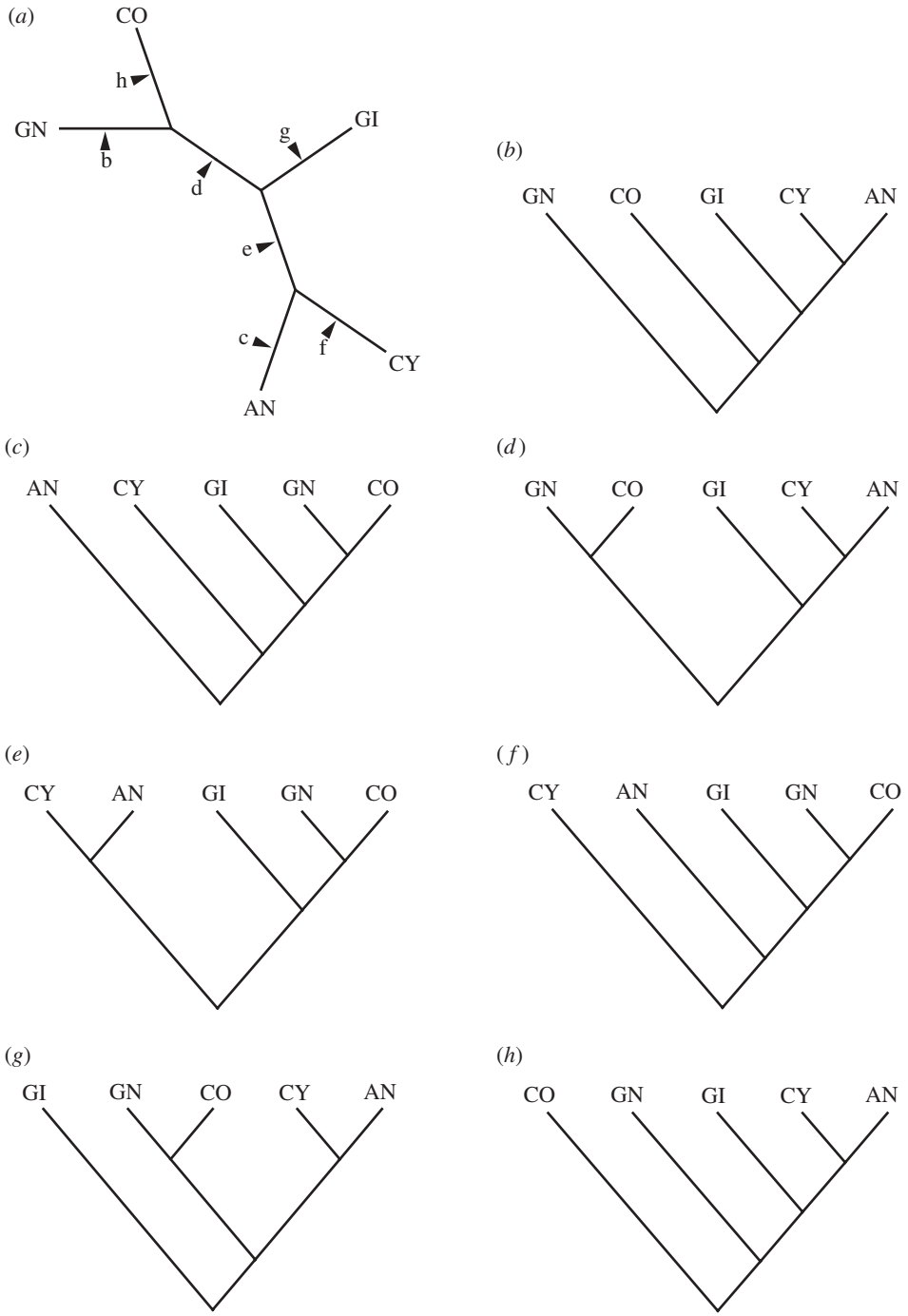

Figure 5. (*a*) Unrooted topology for the rooted tree inferred by Rai *et al.* (2008). (*b*) Rooted topology inferred by Rai *et al.* using ML. (*c*) Rooted topology inferred by Rai *et al.* using MP. (*d–h*) The five additional possible rooted topologies.

topology from such an exercise in the age of phylogenomics? First, it reveals uncertainty regarding the rooting of the seed plant tree that may also exist in other datasets, and this is worth examining. We usually evaluate uncertainty by estimation of bootstrap support values or posterior probabilities. When a clade is highly supported and consistently so across analyses, we rarely test whether the data can reject alternative topologies. This is reasonable enough, except when another line of evidence is contradictory. In the case of seed plants, poorly supported morphological topologies contradict, at least with respect to the rooting of crown seed plants, well-supported molecular topologies. Owing to this disparity in support, the molecular topologies have been favoured. However, morphological trees include many more of the relevant taxa, and the possibility that they are more accurate with respect to the topology of living taxa should be considered. Also, high bootstrap values sometimes support erroneous clades (see Goremykin *et al.* (2003) for a recent case). This clearly happens in seed plant analyses, which have supported *multiple* positions for the gnetophytes, each with bootstrap values of 100 per cent (e.g. Chumley *et al.* 2008; Rai *et al.* 2008). To test the accuracy of the molecular topology that eventually wins out, it may be productive to look at the question in a different way, or by considering other lines of evidence. The duplicate gene-rooting approach, which is a single-gene (protein) analysis in this case, but need not be, is a

Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

different way of looking at the question. In the case of the seed plants, the results do not corroborate results from multi-locus analyses, but they are consistent with results from morphological analyses. This suggests that we should examine the power of other datasets to reject alternative topologies, particularly with respect to the rooting. A second insight from the PHY analyses is that the amino acid data are useful for resolving relationships deep within seed plants, as demonstrated by the parsimonious solution for phytochrome evolution in seed plants. It is notable that slowly evolving nucleotide characters also support the link between cycads and angiosperms but that the amino acid characters do so with stronger support, despite being fewer in number (data not shown; Mathews 2009). While some have argued against the utility of amino acid data relative to nucleotide data (e.g. Simmons *et al.* 2002), it is apparent from the results presented here and in other cases (Hrdy *et al.* 2004; Wildman *et al.* 2007; M. Beilstein, M. Clements & S. Mathews 2009, unpublished data) that they are useful for resolving backbones in a phylogeny. Moreover, it may be more productive to evaluate informativeness on a case-by-case basis (Townsend 2007).

Duplicate gene rooting represents only one option for looking at a phylogenetic question in a different way. It is almost certain that considerable pattern heterogeneity exists in datasets from seed plants. That is, sites are likely to vary not only in their rate of evolution, but also are likely to exhibit distinct patterns of substitution (Pagel & Meade 2004, 2005). For such datasets, the use of mixture models will not only improve the likelihood and better characterize patterns of substitution (manifest as longer trees), but it may also yield different topologies and levels of node support; an empirical example revealed previously undetected uncertainty in some relationships among mammalian orders, whereas it increased confidence in others (Pagel & Meade 2004, 2005). The use of mixture models in a Bayesian context, as advocated by Pagel & Meade (2004, 2005), along with the use of mixed and covarion models that account for heterotachy (e.g. Kolaczkowski & Thornton 2004, 2008), are important tools for exploration of seed plant datasets to verify support and/or reveal uncertainty. Additionally, these datasets warrant exploration with approaches for estimation of species trees from gene trees that avoid the pitfalls associated with concatenation (e.g. Degnan & Rosenberg 2006; Kubatko & Degnan 2007; for a review, see Edwards 2008).

To understand character transitions in seed plants, for example, along the branch leading to angiosperms, a robust hypothesis for all known lineages, living and extinct, is required. Unquestionably, the molecular phylogeny of the major living groups will stabilize in the near future, given the increasing ease with which large numbers of genes from many species can be sequenced (e.g. Cronn *et al.* 2008), and given the advances in analytical approaches that incorporate insights into substitutional processes and the effects of branch-length heterogeneity (e.g. Kolaczkowski & Thornton 2004, 2008; Pagel & Meade 2004, 2005; Matsen & Steel 2007; Edwards 2008). Thus, the areas in which major advances are needed are morphology and integration of insights from morphology and molecules; new fossil data are needed, as are more sophisticated models for evaluating morphological evolution. *Caytonia* is a close relative of angiosperms in some trees (Crane 1985; Hilton & Bateman 2006; Friis *et al.* 2007; Doyle 2008; figure 4), but it represents a very poorly understood fossil gymnosperm (Crane 1985; Taylor & Taylor 2009), making it difficult to devise sound reconstructions of character transitions. Similarly, for corystosperms, which figure in the mostly male theory of angiosperm origin (Frohlich & Parker 2000), there is no whole-plant concept, and their affinities also are uncertain (figure 4). Bennettitales are much better known (Crepet 1972, 1974; Crepet & Delevoryas 1972; Crane & Herendeen 2009), and are consistently placed near angiosperms (Crane 1985; Doyle & Donoghue 1986; Nixon *et al.* 1994; Rothwell & Serbet 1994; Doyle 2006, 2008; Hilton & Bateman 2006; Friis *et al.* 2007; figure 4), but their position relative to other candidate angiosperm sister groups requires further testing. *Archaefructus*, thought by some to be a stem angiosperm (Sun *et al.* 2002), more likely belongs to the water lily clade (Friis *et al.* 2003; Endress & Doyle 2009). In short, much more work on fossils is needed, to produce whole-plant reconstructions for poorly known groups, and to discover the new material that will be necessary for that work and for assessment of homologies. Exciting advances are resulting from the examination of mesofossils and the use of new, non-destructive techniques (Friis *et al.* 2007); equally important will be the discovery and characterization of additional macrofossil material (Crane & Herendeen 2009). More generally, there is cause for optimism that the combination and integration of disparate types of data will be fruitful. First, the quality and quantity of the fossil data continue to improve (e.g. Crane & Herendeen 2009; Friis *et al.* 2009; Stockey & Rothwell 2009), and there is a concurrent push to obtain the comparable data from living taxa that are necessary to improve the context for evaluation of the fossil data. Second, there are more approaches for the exploration of datasets that combine nucleotide or amino acid with morphological characters. A commonly voiced concern with respect to combining sequence and morphological data is that phylogenetic signal in the morphological data will be swamped by signal in the sequence data. At one extreme, because signal in morphological data may be weak, they are viewed as useful only when evaluated in the context of a molecular phylogeny (Scotland *et al.* 2003), but in the case of seed plants, this supposes the availability of a molecular tree that includes many extinct taxa. One of the major obstacles to accurate reconstruction of morphological (and developmental) data on seed plant trees is the lack of information from so many of the relevant groups. This will continue to be a problem in the case of developmental data that cannot be inferred from fossils, and thus, a level of uncertainty will persist around evo-devo scenarios for seed plants. In contrast, morphological reconstructions are expected to improve as fossils become better elucidated. Another approach for

Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

integration that avoids combining sequence and morphological data uses a molecular hypothesis to constrain analyses of morphological data (Doyle 2006, 2008; Hilton & Bateman 2006). The PHY results presented here suggest that in cases where there is uncertainty in the molecular topology, multiple molecular topologies should be used as constraints in the exploration of morphological data. However, it is now possible to move beyond the use of constraints. Mixture models are able to detect and characterize complex historical signals in phylogenetic data, and they can be applied not only to single genes and concatenated alignments, but also to alignments that include morphological characters (Pagel & Meade 2004, 2005). Since amino acid data appear to have less of a tendency to swamp morphological signal (S. Mathews 2009, unpublished data), it will be interesting to explore combinations of morphological characters with amino acids rather than with nucleotides.

The authors are grateful to Ethan Levesque, Mariya Schilz and Kurt Schellenberg for assistance with RACE and TAIL PCR, to Elena Kramer for sharing phytochrome sequences from *Aquilegia*, to Jim Doyle for sharing data and insights and to Michael Donoghue for helping to initiate the duplicate gene-rooting studies of angiosperms and seed plants. This work was supported by NSF grants DEB-0196150 and IBN-021449.

REFERENCES

Abascal, F., Posada, D. & Zardoya, R. 2007 MtArt: a new model of amino acid replacement for Arthropoda. *Mol. Biol. Evol* **24**, 1. (doi:10.1093/molbev/msl136)

Bailey, I. W. 1944 The development of vessels in angiosperms and its significance in morphological research. *Am. J. Bot.* **31**, 421–428. (doi:10.2307/2437302)

Baldauf, S. L., Palmer, J. D. & Doolittle, W. F. 1996 The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl Acad. Sci. USA* **93**, 7749–7754. (doi:10.1073/pnas.93.15.7749)

Barkman, T. J., Chenery, G., McNeal, J. R., Lyons-Weiler, J., Ellisens, W. J., Moore, G., Wolfe, A. D. & dePamphilis, C. W. 2000 Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. *Proc. Natl Acad. Sci. USA* **97**, 13 166–13 171. (doi:10.1073/pnas.220427497)

Bierhorst, D. W. 1971 *Morphology of vascular plants*. New York, NY: Macmillan.

Bowe, L. M., Coat, G. & dePamphilis, C. W. 2000 Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. *Proc. Natl Acad. Sci. USA* **97**, 4092–4097. (doi:10.1073/pnas.97.8.4092)

Brown, J. R. & Doolittle, W. F. 1995 Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl Acad. Sci. USA* **92**, 2441–2445. (doi:10.1073/pnas.92.7.2441)

Burleigh, J. G. & Mathews, S. 2004 Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. *Am. J. Bot.* **91**, 1599–1613. (doi:10.3732/ajb.91.10.1599)

Burleigh, J. G. & Mathews, S. 2007*a* Assessing among-locus variation in the inference of seed plant phylogeny. *Int. J. Plant Sci.* **168**, 111–124. (doi:10.1086/509586)

Burleigh, J. G. & Mathews, S. 2007*b* Assessing systematic error in the inference of seed plant phylogeny. *Int. J. Plant Sci.* **168**, 125–135. (doi:10.1086/509588)

Carlquist, S. 1996 Wood, bark, and stem anatomy of Gnetales: a summary. *Int. J. Plant Sci.* **157**, S58–S76. (doi:10.1086/297404)

Chamberlain, C. J. 1935 *Gymnosperms. Structure and evolution*. Chicago, IL: University of Chicago Press.

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

Chaw, S. M., Parkinson, C. L., Cheng, Y., Vincent, T. M. & Palmer, J. D. 2000 Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc. Natl Acad. Sci. USA* **97**, 4086–4091. (doi:10.1073/pnas.97.8.4086)

Chumley, T. W., McCoy, S. K. R. & Raubeson, L. A. 2008 Gne-deep: exploring Gnetalean affinities in seed plant phylogeny with 83 plastid genes. Botany 2008: Joint Annual Meeting of Canadian Botanical Association, American Fern Society, American Society of Plant Taxonomists, and the Botanical Society of America, Vancouver, British Columbia, Canada. See http://www.2008.botanyconference.org/engine/search/index.php?func=detail&aid=770].

Clack, T., Mathews, S. & Sharrock, R. A. 1994 The phytochrome apoprotein family in *Arabidopsis* is encoded by five genes: the sequences and expression of *PHYD* and *PHYE*. *Plant Mol. Biol.* **25**, 413–427. (doi:10.1007/BF00043870)

Crane, P. R. 1985 Phylogenetic analysis of seed plants and the origin of angiosperms. *Ann. Mo. Bot. Gard.* **72**, 716–793. (doi:10.2307/2399221)

Crane, P. R. & Herendeen, P. S. 2009 Bennettitales from the Grisethorpe Bed (Middle Jurassic) at Cayton Bay, Yorkshire, UK. *Am. J. Bot.* **96**, 284–295. (doi:10.3732/ajb.0800193)

Crepet, W. L. 1972 Investigations of North American cycadeoids: pollination mechanisms in *Cycadeoidea*. *Am. J. Bot.* **59**, 1048–1056. (doi:10.2307/2441490)

Crepet, W. L. 1974 Investigations of North American cycadeoids: the reproductive biology of *Cycadeoidea*. *Palaeontographica B* **148**, 144–169.

Crepet, W. L. 2000 Progress in understanding angiosperm history, success, and relationships: Darwin's abominably 'perplexing phenomenon'. *Proc. Natl Acad. Sci. USA* **97**, 12 939–12 941. (doi:10.1073/pnas.97.24.12939)

Crepet, W. L. & Delevoryas, T. 1972 Investigations of North American cycadeoids: early ovule ontogeny. *Am. J. Bot.* **59**, 209–215. (doi:10.2307/2441403)

Cronn, R., Liston, A., Parks, M., Gernandt, D. S., Shen, R. & Mockler, T. 2008 Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* **36**, e122. (doi:10.1093/nar/gkn502)

Degnan, J. H. & Rosenberg, N. A. 2006 Discordance of species trees with their most likely gene trees. *PLoS Genet.* **2**, e68. (doi:10.1371/journal.pgen.0020068)

de la Torre, J., Egan, M., Katari, M., Brenner, E., Stevenson, D., Coruzzi, G. & DeSalle, R. 2006 ESTimating plant phylogeny: lessons from partitioning. *BMC Evol. Biol.* **6**, 48. (doi:10.1186/1471-2148-6-48)

Donoghue, M. J. & Doyle, J. A. 2000 Seed plant phylogeny: demise of the anthophyte hypothesis? *Curr. Biol.* **10**, R106–R109. (doi:10.1016/S0960-9822(00)00304-3)

Donoghue, M. J. & Mathews, S. 1998 Duplicate genes and the root of angiosperms, with an example using

Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

phytochrome sequences. *Mol. Phylogen. Evol.* **9**, 489–500. (doi:10.1006/mpev.1998.0511)
Doyle, J. A. 1996 Seed plant phylogeny and the relationships of Gnetales. *Int. J. Plant Sci.* **157**, S3–S39. (doi:10.1086/297401)
Doyle, J. A. 2006 Seed ferns and the origin of angiosperms. *J. Torrey Bot. Soc.* **133**, 169–209.
Doyle, J. A. 2008 Integrating molecular phylogenetic evidence and paleobotanical evidence on the origin of the flower. *Int. J. Plant Sci.* **167**, 816–843.
Doyle, J. A. & Donoghue, M. J. 1986 Seed plant phylogeny and the origin of angiosperms: an experimental cladistic approach. *Bot. Rev.* **52**, 321–431. (doi:10.1007/BF02861082)
Doyle, J. A. & Donoghue, M. J. 1987 The origin of angiosperms: a cladistic approach. In *The origins of angiosperms and their biological consequences* (eds E. M. Friis & W. G. Chalconer & P. R. Crane), pp. 17–49. Cambridge, UK: Cambridge University Press.
Edwards, S. V. 2008 Is a new and general theory of molecular systematics emerging? *Evolution* **63**, 1–19. (doi:10.1111/j.1558-5646.2008.00549.x)
Endress, P. K. & Doyle, J. A. 2009 Reconstructing the ancestral angiosperm flower and its initial specializations. *Am. J. Bot.* **96**, 22–66. (doi:10.3732/ajb.0800047)
Felsenstein, J. 1978 Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410. (doi:10.2307/2412923)
Felsenstein, J. 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376. (doi:10.1007/BF01734359)
Friis, E. M., Chalconer, W. G. & Crane, P. R. 1987 *The origins of angiosperms and their biological consequences*. Cambridge, UK: Cambridge University Press.
Friis, E. M., Doyle, J. A., Endress, P. K. & Leng, Q. 2003 Archaefructus–angiosperm precursor or specialized early angiosperm? *Trends Plant Sci.* **8**, 369–373. (doi:10.1016/S1360-1385(03)00161-4)
Friis, E. M., Crane, P. R., Pedersen, K. R., Bengtson, S., Donoghue, P. C. J., Grimm, G. W. & Stampanoni, M. 2007 Phase-contrast X-ray microtomography links Cretaceous seeds with Gnetales and Bennettitales. *Nature* **450**, 549–552. (doi:10.1038/nature06278)
Friis, E. M., Pedersen, K. R. & Crane, P. R. 2009 Early Cretaceous mesofossils from Portugal and eastern North America related to the Bennettitales–Erdtmanithecales–Gnetales group. *Am. J. Bot.* **96**, 252–283. (doi:10.3732/ajb.0800113)
Frohlich, M. W. & Parker, D. S. 2000 The mostly male theory of flower evolutionary origins: from genes to fossils. *Syst. Bot.* **25**, 155–170. (doi:10.2307/2666635)
Frohman, M. A., Dush, M. K. & Martin, G. R. 1988 Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl Acad. Sci. USA* **85**, 8998–9002. (doi:10.1073/pnas.85.23.8998)
Gogarten, J. P. *et al.* 1989 Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl Acad. Sci. USA* **86**, 6661–6665. (doi:10.1073/pnas.86.17.6661)
Goremykin, V. V., Hirsch-Ernst, K. I., Wolfl, S. & Hellwig, F. H. 2003 Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Mol. Biol. Evol.* **20**, 1499–1505. (doi:10.1093/molbev/msg159)
Graham, S. W. & Iles, W. J. D. 2009 Different gymnosperm outgroups have (mostly) congruent signal regarding the root of flowering plant phylogeny. *Am. J. Bot.* **96**, 216–227. (doi:10.3732/ajb.0800320)
Graham, S. W. & Olmstead, R. G. 2000 Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *Am. J. Bot.* **87**, 1712–1730. (doi:10.2307/2656749)
Gugerli, F., Sperisen, C., Buchler, U., Brunner, I., Brodbeck, S., Palmer, J. D. & Qiu, Y. L. 2001 The evolutionary split of Pinaceae from other conifers: evidence from an intron loss and a multigene phylogeny. *Mol. Phylogen. Evol.* **21**, 167–175. (doi:10.1006/mpev.2001.1004)
Hajibabaei, M., Xia, J. & Drouin, G. 2006 Seed plant phylogeny: gnetophytes are derived conifers and a sister group to Pinaceae. *Mol. Phylogen. Evol.* **40**, 208–217. (doi:10.1016/j.ympev.2006.03.006)
Hamby, R. K. & Zimmer, E. A. 1992 Ribosomal RNA as a phylogenetic tool in plant systematics. In *Molecular systematics of plants* (eds P. S. Soltis, D. E. Soltis & J. J. Doyle), pp. 50–91. New York, NY: Chapman and Hall.
Hilton, J. & Bateman, R. M. 2006 Pteridosperms are the backbone of seed-plant phylogeny. *J. Torrey Bot. Soc.* **133**, 119–168.
Hrdy, I., Hirt, R. P., Dolezal, P., Bardonova, L., Foster, P. G., Tachezy, J. & Martin Embley, T. 2004 Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* **432**, 618–622. (doi:10.1038/nature03149)
Huelsenbeck, J. P. 1998 Systematic bias in phylogenetic analysis: is the Strepsiptera problem solved? *Syst. Biol.* **47**, 519–537.
Huelsenbeck, J. P. & Ronquist, F. 2001 MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755. (doi:10.1093/bioinformatics/17.8.754)
Huelsenbeck, J. P., Hillis, D. M. & Jones, R. 1996 Parametric bootstrapping in phylogenetics: applications and performance. In *Molecular zoology: Advances, strategies, and protocols* (eds J. D. Ferraris & S. R. Palumbi), pp. 19–45. New York, NY: Wiley-Liss.
Huelsenbeck, J. P., Joyce, P., Lakner, C. & Ronquist, F. 2008 Bayesian analysis of amino acid substitution models. *Phil. Trans. R. Soc. B* **363**, 3941–3953. (doi:10.1098/rstb.2008.0175)
Iwabe, N., Kuma, K.-I., Hasegawa, M., Osawa, S. & Miyata, T. 1989 Evolutionary relationship of Archaebacteria, Eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl Acad. Sci. USA* **86**, 9355–9359. (doi:10.1073/pnas.86.23.9355)
Kolaczkowski, B. & Thornton, J. W. 2004 Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**, 980–984. (doi:10.1038/nature02917)
Kolaczkowski, B. & Thornton, J. W. 2008 A mixed branch length model of heterotachy improves phylogenetic accuracy. *Mol. Biol. Evol.* **25**, 1054–1066. (doi:10.1093/molbev/msn042)
Kubatko, L. S. & Degnan, J. H. 2007 Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* **56**, 17–24. (doi:10.1080/10635150601146041)
Lake, J. A., Skophammer, R. G., Herbold, C. W. & Servin, J. A. 2009 Genome beginnings: rooting the tree of life. *Phil. Trans. R. Soc. B* **364**, 2177–2185. (doi:10.1098/rstb.2009.0035)
Le, S. Q. & Gascuel, O. 2008 An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320. (doi:10.1093/molbev/msn067)
Liu, Y.-G. & Chen, Y. 2007 High-efficiency thermal asymmetric interlaced PCR for amplification of unknown flanking sequences. *BioTechniques* **43**, 649–656. (doi:10.2144/000112601)
Loconte, H. & Stevenson, D. W. 1990 Cladistics of the Spermatophyta. *Brittonia* **42**, 197–211. (doi:10.2307/2807216)

Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

Magallón, S. Submitted. Using fossils to break long branches in molecular dating: a comparison of relaxed clocks applied to the origin of angiosperms.

Magallón, S. & Sanderson, M. J. 2002 Relationships among seed plants inferred from highly conserved genes: sorting conflicting phylogenetic signals among ancient lineages. *Am. J. Bot.* **89**, 1991–2006. (doi:10.3732/ajb.89.12.1991)

Magallón, S. A. & Sanderson, M. J. 2005 Angiosperm divergence times: the effect of genes, codon positions, and time constraints. *Evolution* **59**, 1653–1670. (doi:10.1554/04-565.1)

Mathews, S. 2006 Phytochrome-mediated development in land plants: red light sensing evolves to meet the challenges of changing light environments. *Mol. Ecol.* **15**, 3483–3503. (doi:10.1111/j.1365-294X.2006.03051.x)

Mathews, S. 2009 Phylogenetic relationships among seed plants: persistent questions and the limits of molecular data. *Am. J. Bot.* **96**, 228–236. (doi:10.3732/ajb.0800178)

Mathews, S. & Donoghue, M. J. 1999 The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* **286**, 947–950. (doi:10.1126/science.286.5441.947)

Matsen, F. A. & Steel, M. 2007 Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Syst. Biol.* **56**, 767–775. (doi:10.1080/10635150701627304)

McCoy, S. R., Kuehl, J. V., Boore, J. L. & Raubeson, L. A. 2008 The complete plastid genome sequence of *Welwitschia mirabilis*: an unusually compact plastome with accelerated divergence rates. *BMC Evol. Biol.* **8**, 103.

Nickrent, D. L., Parkinson, C. L., Palmer, J. D. & Duff, R. J. 2000 Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol. Biol. Evol.* **17**, 1885–1895.

Nixon, K. C., Crepet, W. L., Stevenson, D. & Friis, E. M. 1994 A reevaluation of seed plant phylogeny. *Ann. Mo. Bot. Gard.* **81**, 484–533. (doi:10.2307/2399901)

Pagel, M. & Meade, A. 2004 A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.* **53**, 571–581.

Pagel, M. & Meade, A. 2005 Mixture models in phylogenetic inference. In *Mathematics of evolution and phylogeny* (ed. O. Gascuel), pp. 121–142. Oxford, UK: Oxford University Press.

Paradis, E., Claude, J. & Strimmer, K. 2004 APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290. (doi:10.1093/bioinformatics/btg412)

Parenti, L. R. 1980 A phylogenetic analysis of the land plants. *Biol. J. Linn. Soc.* **13**, 225–242. (doi:10.1111/j.1095-8312.1980.tb00084.x)

Parkinson, C. L., Adams, K. L. & Palmer, J. D. 1999 Multigene analyses identify the three earliest lineages of extant flowering plants. *Curr. Biol.* **9**, 1485–1488. (doi:10.1016/S0960-9822(00)80119-0)

Qiu, Y.-L. *et al.* 1999 The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* **402**, 404–407. (doi:10.1038/46536)

R Development Core Team. 2009 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. See http://www.R-project.org.

Rai, H. S., Reeves, P. A., Peakall, R., Olmstead, R. G. & Graham, S. W. 2008 Inference of higher-order conifer relationships from a multi-locus plastid data set. *Can. J. Bot.* **86**, 658–669. (doi:10.1139/B08-062)

Rambaut, A. & Drummond, A. J. 2007 *MCMC trace analysis tool version v1.4, 2003–2007*. See http://tree.bio.ed.ac.uk/software/tracer/.

Ronquist, F. & Huelsenbeck, J. P. 2003 MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574. (doi:10.1093/bioinformatics/btg180)

Rothwell, G. W. & Serbet, R. 1994 Lignophyte phylogeny and the evolution of Spermatophytes: a numerical cladistic analysis. *Syst. Bot.* **19**, 443–482. (doi:10.2307/2419767)

Rydin, C. & Källersjö, M. 2002 Taxon sampling and seed plant phylogeny. *Cladistics* **18**, 484–513.

Rydin, C., Källersjö, M. & Friis, E. M. 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. (doi:10.1086/338321)

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

Schliep, K. 2009 phangorn: phylogenetic analysis in R. R package version 0.99-2. http://CRAN.R-project.org/package=phangorn.

Schmidt, M. & Schneider-Poetsch, H. A. 2002 The evolution of gymnosperms redrawn by phytochrome genes: the *Gnetatae* appear at the base of the gymnosperms. *J. Mol. Evol.* **54**, 715–724. (doi:10.1007/s00239-001-0042-9)

Scotland, R. W., Olmstead, R. G. & Bennett, J. R. 2003 Phylogeny reconstruction: the role of morphology. *Syst. Biol.* **52**, 539–548.

Sharrock, R. A. & Quail, P. H. 1989 Novel phytochrome sequences in *Arabidopsis thaliana*: structure, evolution, and differential expression of a plant regulatory photoreceptor family. *Genes Dev.* **3**, 1745–1757. (doi:10.1101/gad.3.11.1745)

Shimodaira, H. 2002 An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508. (doi:10.1080/10635150290069913)

Shimodaira, H. 2008 Testing regions with nonsmooth boundaries via multiscale bootstrap. *J. Statist. Plann. Inf.* **138**, 1227–1241. (doi:10.1016/j.jspi.2007.04.001)

Simmons, M. P., Ochoterena, H. & Freudenstein, J. V. 2002 Amino acid vs. nucleotide characters: challenging preconceived notions. *Mol. Phylogen. Evol.* **24**, 78–90. (doi:10.1016/S1055-7903(02)00202-6)

Slowinski, J. B., Knight, A. & Rooney, A. P. 1997 Inferring species trees from gene trees: a phylogenetic analysis of the Elapidae (Serpentes) based on the amino acid sequences of venom proteins. *Mol. Phylogen. Evol.* **8**, 349–362. (doi:10.1006/mpev.1997.0434)

Soltis, D. E. *et al.* 2000 Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Bot. J. Linn. Soc.* **133**, 381–461.

Soltis, D. E., Soltis, P. S. & Zanis, M. J. 2002 Phylogeny of seed plants based on evidence from eight genes. *Am. J. Bot.* **89**, 1670–1681. (doi:10.3732/ajb.89.10.1670)

Stamatakis, A. 2006 RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690. (doi:10.1093/bioinformatics/btl446)

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

Stefanović, S., Jager, M., Deutsch, J., Broutin, J. & Masselot, M. 1998 Phylogenetic relationships of conifers inferred from partial 28S rRNA gene sequences. *Am. J. Bot.* **85**, 688–697. (doi:10.2307/2446539)

Stockey, R. A. & Rothwell, G. W. 2009 Distinguishing angiophytes from the earliest angiosperms: a Lower Cretaceous (Valanginian–Hauterivian) fruit-like reproductive

Downloaded from rstb.royalsocietypublishing.org on May 11, 2012

structure. *Am. J. Bot.* **96**, 323–335. (doi:10.3732/ajb.0800295)

Stockey, R. A., Graham, S. W. & Crane, P. R. 2009 Introduction to the Darwin special issue: the abominable mystery. *Am. J. Bot.* **96**, 3–4. (doi:10.3732/ajb.0800402)

Sun, G., Ji, Q., Dilcher, D. L., Zheng, S., Nixon, K. C. & Wang, X. 2002 Archaefructaceae, a new basal angiosperm family. *Science* **296**, 899–904. (doi:10.1126/science.1069439)

Swofford, D. L. 2002 *PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4.0b10*. Sunderland, MA: Sinauer Associates.

Taylor, E. L. & Taylor, T. N. 2009 Seed ferns from the late Paleozoic and Mesozoic: any angiosperm ancestors lurking there? *Am. J. Bot.* **96**, 237–251. (doi:10.3732/ajb.0800202)

Townsend, J. P. 2007 Profiling phylogenetic informativeness. *Syst. Biol.* **56**, 222–231. (doi:10.1080/10635150701311362)

Whelan, S. 2008 The genetic code can cause systematic bias in simple phylogenetic models. *Phil. Trans. R. Soc. B* **363**, 4003–4011. (doi:10.1098/rstb.2008.0171)

Whelan, S. & Goldman, N. 2001 A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699.

Wildman, D. E. *et al.* 2007 Genomics, biogeography, and the diversification of placental mammals. *Proc. Natl Acad. Sci. USA* **104**, 14 395–14 400. (doi:10.1073/pnas.0704342104)

Wu, C.-S., Wang, Y.-N., Liu, S.-M. & Chaw, M. 2007 Chloroplast genome (cpDNA) of *Cycas taitungensis* and 56 cp protein-coding genes of *Gnetum parvifolium*: insights into cpDNA evolution and phylogeny of extant seed plants. *Mol. Biol. Evol.* **24**, 1366–1379. (doi:10.1093/molbev/msm059)

Yang, Z. 1993 Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**, 1396–1401.

Zanis, M. J., Soltis, D. E., Soltis, P. S., Mathews, S. & Donoghue, M. J. 2002 The root of the angiosperms revisited. *Proc. Natl Acad. Sci. USA* **99**, 6848–6853. (doi:10.1073/pnas.092136399)