Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution

William Martin[1], Oliver Deusch[1], Nadine Stawski[1], Nicole Grünheit[1] and Vadim Goremykin[2]

[1]Institute of Botany III, University of Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany
[2]Institut für Spezielle Botanik, Universität Jena, Philosophenweg 16, D-07743 Jena, Germany

The traditional approach to plant molecular phylogenetics involves amplifying, sequencing and analyzing one or a few genes from many species and is conducive to broad taxon sampling. An independent approach involves chloroplast genome sequencing, providing much larger amounts of data per taxon but for a smaller number of species. In principle, the two strategies can inform each other but in practice their results sometimes conflict for reasons that are currently debated. An Opinion article published in the October 2004 issue of *Trends in Plant Science* cautioned against the pursuit of genome-based phylogenies. Here, we provide a different perspective on issues at the heart of the current debate and defend the use of chloroplast genome phylogenetics for crucial species because it provides an independent test of hypotheses generated by the traditional approach.

Taxon sampling is not the only problem we face

In an Opinion article published in the October 2004 issue of *Trends in Plant Science*, Douglas Soltis *et al.* [1] cautioned against the use of data from complete genomes for studying evolution. In the main, their arguments centered on the premise that taxon sampling (investigating many lineages) is, in general, far more important than site sampling (investigating many sites from many genes from a few crucial lineages) in building trees from sequence data. Their case was argued with examples from yeast phylogeny, mitochondrial genome phylogeny and – of specific relevance to the plant community – a question regarding the most primitive angiosperms, in particular, the position of *Amborella*. Various issues were considered by Soltis *et al.* [1], including the presumed utility or suspected vagary of including highly variable third codon positions in investigations of deeper phylogenetic relationships. However, their salient argument was that studying many genes from the most crucial species is not advisable, whereas studying a few genes from many species is. They conclude with the advice: 'As scientists assemble the tree of life, perhaps we need to rethink the strategies behind some ongoing projects. Some funded initiatives are primarily or exclusively using whole organellar genome sequencing for a small number of taxa. Our example of *Amborella* indicates that such a strategy can be seriously flawed…' (Ref. [1], p. 482).

Viewing matters from a different standpoint, we address three issues here. First, we argue that the jury is still out concerning the position of *Amborella* in angiosperm phylogeny, hence the critique leveled by Soltis *et al.* [1] – that results recently obtained by Vadim Goremykin *et al.* [2,3] from complete chloroplast genome phylogeny are artefactual – is premature. Second, we point out that recent findings show that the processes of sequence evolution as they occur in nature deviate substantially from the processes assumed by current phylogeny inference programs. Such deviations include lineage-specific departures from an assumed common distribution of across-site rate variation (covarion evolution) and lineage-specific departures from an assumed symmetric substitution model (compositional heterogeneity). As with taxon sampling, these processes are crucial to phylogenetic inference but, in contrast with taxon sampling, they are often left unmentioned [1]. Finally, we argue that genome phylogenies are not something against which to caution, rather they offer a unique opportunity to understand plant evolution better and are an independent test of hypotheses generated by traditional means.

Position of Amborella

Amborella is a dicot plant and has attracted the interest of the plant evolutionary community because molecular phylogenies, based upon the analyses of one or a few genes [4–6], have surprisingly suggested that it might represent the most basal angiosperm lineage, although independent analyses have challenged that view [7,8]. By contrast, analyses of the complete *Amborella* chloroplast genome, encompassing 61 genes and many thousands of sites, unexpectedly, did not confirm that basal position [2,3]. Instead, these analyses suggest that monocots, albeit represented by only a few grass lineages, assume a more basal position in angiosperm phylogeny than *Amborella* does. Soltis *et al.* [1] argue that the results of Goremykin *et al.* [2,3] are an artefact of sparse taxon sampling.

We do not doubt that taxon sampling is important in phylogenetics, although the issues of just how important it

Corresponding author: Martin, W. (w.martin@uni-duesseldorf.de).
Available online 2 April 2005

 1360-1385/$ - see front matter doi:10.1016/j.tplants.2005.03.007

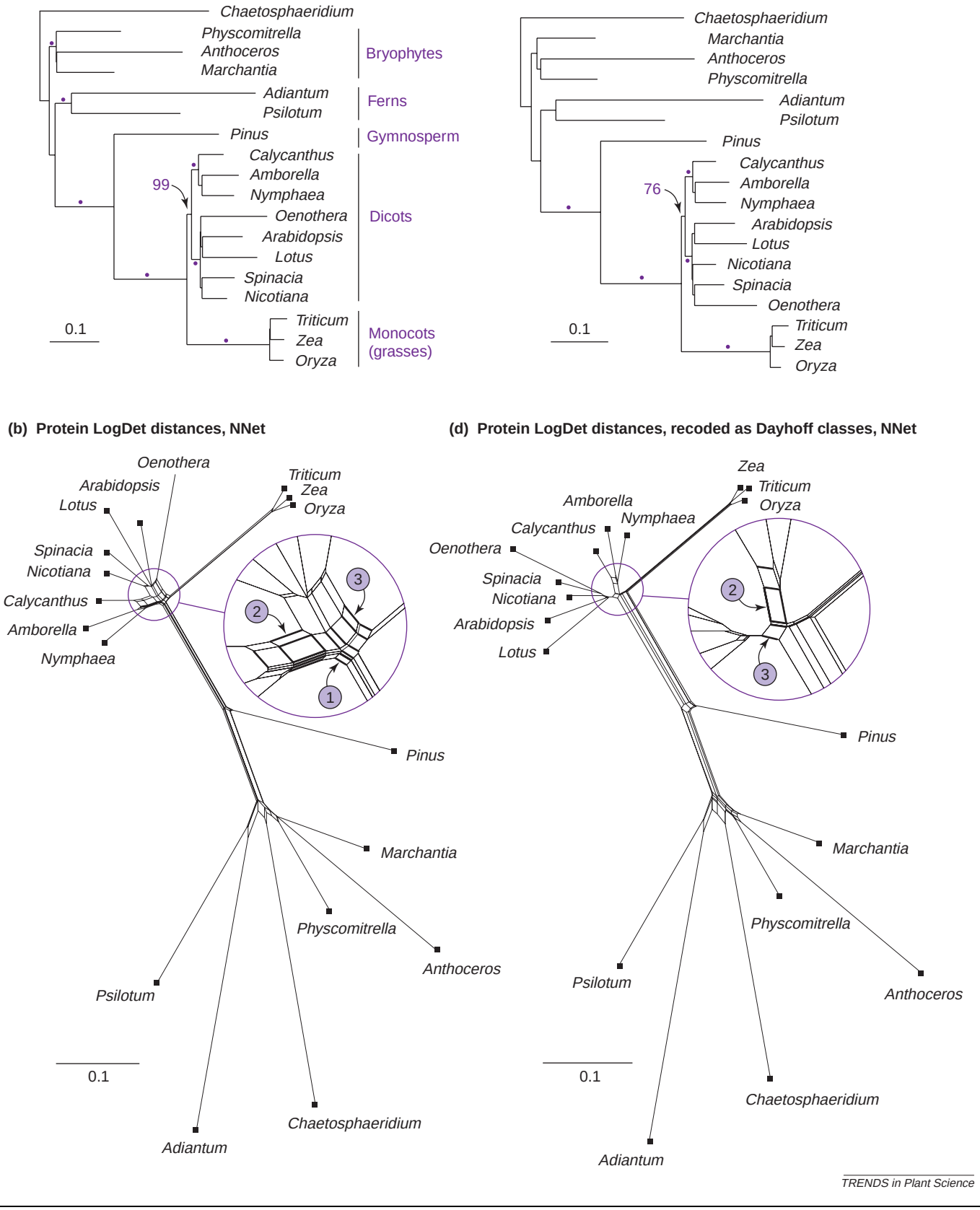

Figure 1. Neighbor-Joining (NJ) trees and NeighborNet (NNet) splits graphs for 18 taxa using the 61 proteins from chloroplast genomes used previously [2]. Proteins were aligned individually with ClustalW and concatenated. The initial alignment contained 13 556 sites per genome total including 1369 gapped sites that were excluded from analysis, leaving 12 187 amino acid sites per genome for log determinant (LogDet) distance estimates with estimation and removal of invariant sites using the program LDDist [21]. From this, NJ trees [39] and NNet splits graphs [20] were constructed; NNet splits graphs were visualized with Splitstree [40]. **(a)** NJ tree without recoding of amino acids. Purple dots indicate branches with bootstrap proportions (BP) ≥95%; the value for the branch placing monocots deep is indicated. **(b)** NNet without recoding of

is for phylogenies and the exact reasons why it is important are not yet resolved [9,10]. Indeed, Soltis *et al.* [1] concede that '…what constitutes "adequate" taxon sampling is not always a straightforward issue'. Goremykin *et al.* [2,3] sequenced the chloroplast genomes of *Amborella* and *Nymphaea*, another angiosperm suspected to be of basal lineage, in the expectation of finding support for their basal positions as the earlier studies had suggested. They employed various phylogenetic methods, including many not mentioned by Soltis *et al.* [1], and found the contrary result. Goremykin *et al.* [2,3] emphasized that increased taxon sampling in the future would be important to test those findings.

Saša Stefanović *et al.* [11] recently improved the taxon sampling for 40 of 61 genes of the chloroplast dataset to address the position of *Amborella* by including data from the monocot *Acorus*. Their results sometimes indicated a basal position for *Amborella* in analyses where *Acorus* was substituted for, rather than added to, the grasses. However, when the taxon sampling was increased by including *Acorus* and the grasses, a basal position for either grasses or *Amborella* was obtained, depending upon the phylogenetic method used [11], whereby 40-gene analyses, including *Acorus*, grasses, *Amborella* and *Nymphaea*, were not reported [11]. At face value, this would suggest that the method of phylogenetic inference used and the underlying assumptions of the models involved [12,13] have at least as much influence on the inferred position of *Amborella* as increasing taxon sampling among crucial lineages does. A recent supertree analysis encompassing more genes and broader taxon sampling among higher plants than reported by either Goremykin *et al.* [2,3] or Soltis *et al.* [1] placed a monocot, not *Amborella*, at the base of the angiosperm tree [14].

Compositional heterogeneity

One aspect of data analysis deemed to be of particular importance by Soltis *et al.* [1] was homoplasy, that is, sites that have the same nucleotide or amino acid state owing to parallel substitutions. One of the most common causes of homoplasy is compositional heterogeneity (compositional bias). When large numbers of substitutions have occurred, directional (biased) substitution processes in independent lineages can result in elevated GC- or AT-content throughout the gene or genome [15]. Current phylogenetic methods, which assume a symmetric substitution model, cannot easily deal with such asymmetric (biased) substitution processes [16]. AT content is extremely high in chloroplast genomes, with codon third positions possessing >85% AT in some lineages. Although Soltis *et al.* [1] argue that including such third positions improves phylogenetic results, such extreme bias led to the decision to exclude third codon position data in the analyses of Goremykin *et al.* [2,3]. Compositional bias can even surface at the amino acid level, affecting the amino acid composition of protein sequences [16]. Compositional bias is known to have a strong influence on phylogenetic results, both in single-gene and in genome trees [17], but its affects can be compensated for by log determinant (LogDet) distance estimates [15,18], also known as paralinear distances [19]. When we examined chloroplast genome phylogeny with protein LogDet distances, we found that grasses, not *Amborella*, are basal among angiosperms. The result, expressed as a bifurcating tree, shows a basal position for the grasses (Figure 1a) with a high bootstrap proportion (99%) that is hardly surprising for long sequences [17].

But are trees always the best way to represent molecular data? Molecular data usually contain conflicting signals, some of which might be the trace of evolutionary history [18]. When there are conflicting signals with regard to a particular branch in the data, trees will tend to show the stronger signal detected with the given model, even if it is only slightly stronger, and depict the sum of signals as a system of compatible splits. Network approaches to sequence analysis such as NeighborNet (NNet) [20] can also represent incompatible splits too, not just the ones that will fit onto a bifurcating tree. To see if there was a competing signal in the data placing *Amborella* at the base of the angiosperms rather than among the other dicots, we examined the LogDet distances [21] among the chloroplast sequences with NeighborNet and found no split that would place *Amborella* or other members of the 'ANITA' group suspected of early-branching, which includes *Amborella* and *Nymphaea*, basal to the remaining angiosperms (Figure 1b). A strong split uniting members of the ANITA group with *Calycanthus* was detected, albeit alongside a conflicting split grouping *Calycanthus* with the other dicots (Figure 1b).

Another way to deal with homoplasy in data, whether owing to multiple substitutions or to bias, is to recode the data into more general classes of substitutions. With nucleotide sequences, this is usually achieved by recoding four-state GACT data into pyrimidines and purines (RY-coding). This approach can be useful in detecting more-ancient signals in data (at the expense of resolution at the tips of the tree) for the purpose of recovering deeper relationships [22]. Recoding procedures for amino acids have also been explored to some extent to recover ancient phylogenetic signals [23,24]. We therefore recoded the chloroplast amino acid data into the six Dayhoff groups used by Ivan Hrdy *et al.* [23] and repeated the LogDet analysis. In the bifurcating tree (Figure 1c), the bootstrap proportion for the basal position of grasses was lower but, again, in the network, no split was observed that would indicate the basal position of *Amborella* (Figure 1d). Adding more outgroups to the angiosperm sample did not change that result (Figure 2).

Although it might appear as if the chloroplast genome NNet uncovers a large amount of conflicting signals (Figures 1 and 2), the data are tree-like. For comparison, an NNet of LogDet distances for the five-gene nucleotide

amino acids. The inset highlights the split putting grasses basal (1), that uniting *Calycanthus, Nymphaea* and *Amborella* (2), and a conflicting split uniting *Calycanthus* with other dicots (3). **(c)** NJ tree with recoding into the Dayhoff classes used by Hrdy *et al.* [40]: C, STPAG, NDEQ, HRK, MILV and FYW. **(d)** NNet with recoding of amino acids into Dayhoff classes. The inset highlights the split uniting *Calycanthus, Nymphaea* and *Amborella* (2), and a conflicting split uniting *Calycanthus* with other dicots (3). The scale bar indicates a distance of 0.1 substitutions per site. Alignment available upon request.

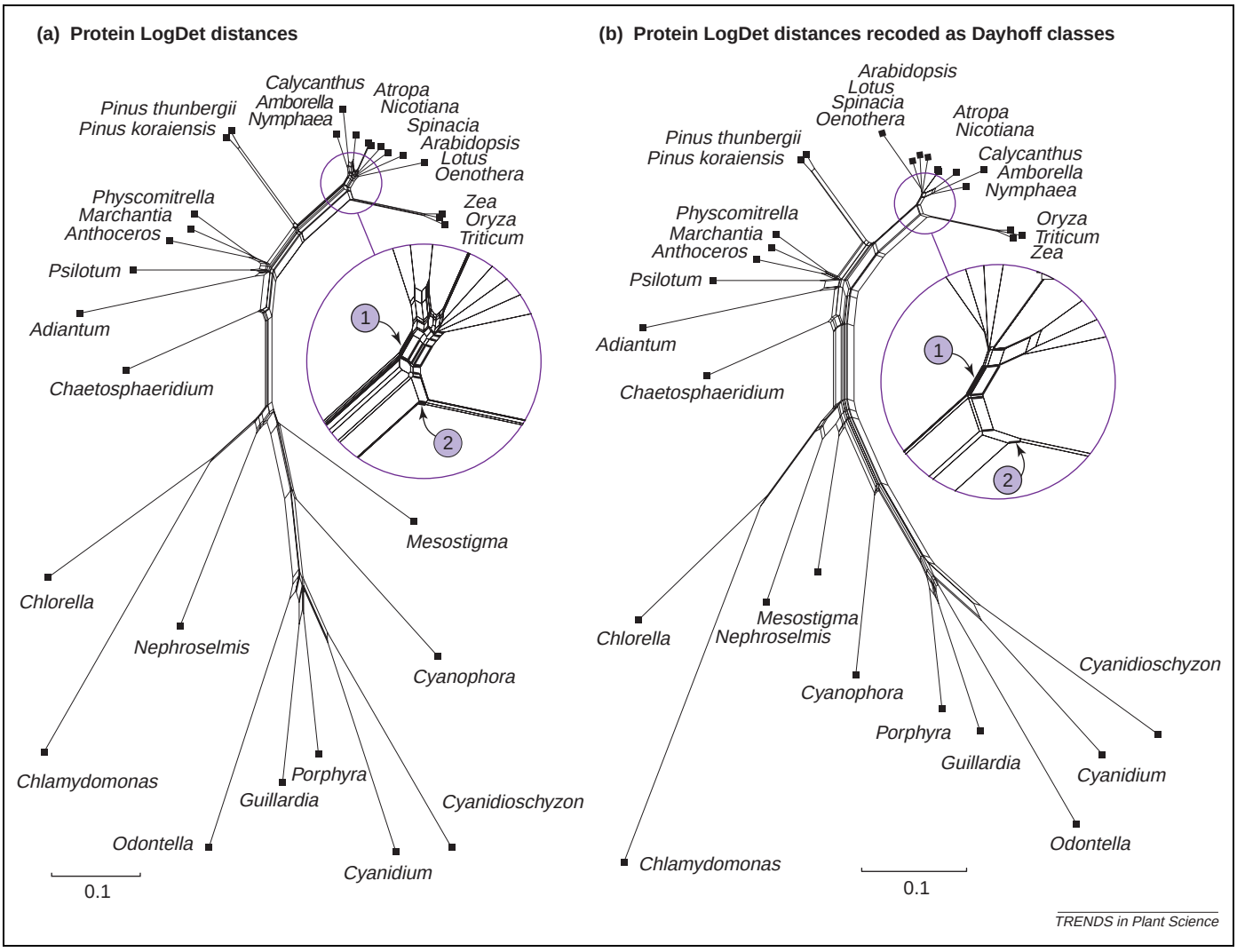

Figure 2. NeighborNet (NNet) splits graphs using more outgroups (30 taxa total) for 43 protein coding sequences from chloroplast genomes. The initial concatenated alignment contained 16 100 sites per genome total including 6644 gapped sites that were excluded from analysis, leaving 9456 amino acid sites per genome for log determinant (LogDet) distance estimates with estimation and removal of invariant sites using LDDist [21], from which NNet planar graphs [20] were constructed and visualized with Splitstree [40]. **(a)** Without recoding of amino acids. The inset highlights the split putting grasses basal (1) and a conflicting split uniting grasses with *Oenothera* (2). **(b)** With recoding as Dayhoff classes. The inset highlights the split putting grasses basal (1) and a conflicting split uniting grasses with *Calycanthus, Nymphaea* and *Amborella* (2). The scale bar indicates a distance of 0.1 substitutions per site. Alignment available upon request.

sequence data of Yin-Long Qiu *et al.* [5], who reported a basal position for *Amborella*, is shown in Figure 3. Notably, NNet detects a split placing *Amborella* and Nymphaeales at the base of the angiosperms in that data, in agreement with the findings of Qiu *et al.* [5], but there are numerous conflicting signals in the data as well. Finding the correct placement of the outgroup root among the rather short internal edges linking major angiosperm lineages (Figures 1–3) is a challenging phylogenetic problem [12,13].

Covarions

Different tree-building methods place *Amborella* either basal among angiosperms or with other dicots on the basis of chloroplast genome data [1–8,11], which draws attention to another important unresolved issue not mentioned by Soltis *et al.* [1] – whether the substitution models currently used in phylogenetic methods approximate the true underlying evolutionary process of sequence change sufficiently to yield results that reflect evolutionary history (rather than reflecting numerical processes within a computer). Possible concerns here are not only asymmetric substitution processes resulting in compositional bias [15–19] as discussed above, but also changes in the proportions of sites that are free to vary in different lineages [25]. The substitution models currently used allow for the possibility that substitution rates can vary across sites but they do not allow sites to alter their rate class in different lineages [26]. However, there is increasing evidence that sites do undergo rate class changes in lineage-specific manners during evolution. This mode of evolutionary change is generally referred to as covarion evolution and it has been observed in sequences derived from eukaryotes, prokaryotes and organelles [25–34]. Some patterns of covarion evolution can lead to current phylogenetic methods strongly supporting an incorrect tree [25,32,33].

Covarion patterns of sequence change have recently been found to exist in chloroplast gene data too [34]. Although yet to be investigated in detail, the conflicting

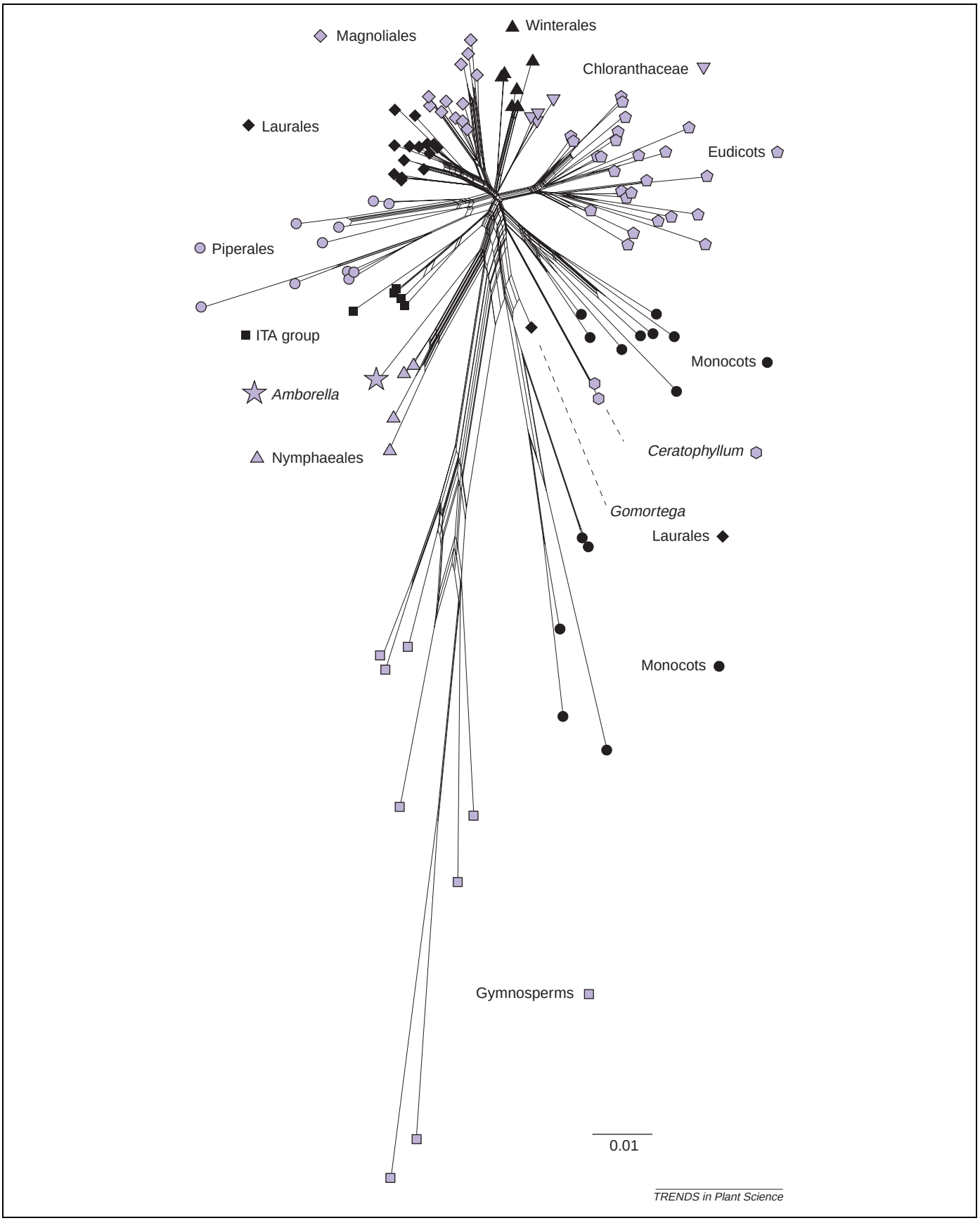

Figure 3. NeighborNet (NNet) splits graph for the five-gene dataset from Qiu *et al.* [5]. The alignment (see http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v402/n6760/abs/402404a0_fs.html&dynoptions=doi1111006742) contained 8733 sites per taxon for nucleotide log determinant (LogDet) distance estimates with LDDist [20], from which NNet splits graphs [18] were constructed and visualized with Splitstree [41]. The major groups indicated with symbols at each leaf correspond 1:1 to those indicated in Figure 1 of Ref. [5]. In Ref. [5], monocots and Laurales were monophyletic; in the present analysis, five monocots (*Potamogeton*, *Triglochin*, *Alisma*, *Acorus*1 and *Acorus*2, in counterclockwise order) and one member of the Laurales (*Gomortega*) did not share a split uniting them with the remainder of their corresponding assemblages. The scale bar indicates a distance of 0.1 substitutions per site.

results of Goremykin *et al.* [2,3], Soltis *et al.* [1], and other studies reporting different positions for *Amborella* [1–8,11], might all be affected by this form of model misspecification. A characteristic feature of covarion evolution is the presence of sites in a multiple sequence alignment where substitutions are not accepted in one group of taxa but where they are readily accepted in another group of taxa [29,30]. This indicates lineage-specific differences in rate class for such sites across the tree in violation of the assumptions regarding sequence evolution phylogeny inference programs used currently. Such covarion patterns might, for example, reflect changes in functional or structural constraints for the same protein in different lineages [34]. However, regardless of the underlying biochemical cause(s), covarion patterns pose severe problems in current phylogeny inference models, allowing sites to have different rates but not allowing sites to change their rate class in different lineages [25–34].

Unexpected twist

An issue of an altogether different nature has recently arisen that complicates the *Amborella* debate in an unexpected manner, underscoring the need to have complete organelle genome sequences for comparison. Using the traditional PCR approach, Ulfar Bergthorsson *et al.* [35] recently amplified multiple copies of *Amborella* mitochondrial-related sequences, some of which branch basal among angiosperms, as Soltis *et al.* [1] have found, some of which branch with dicots, as Goremykin *et al.* have found [2,3], and some of which branch with mosses, which nobody else has previously found. Bergthorsson *et al.* [35] interpret this finding as evidence for massive horizontal gene transfer to the *Amborella* lineage from a myriad of different donor plant lineages. Notably, the *Amborella* chloroplast genome, the only complete-contig genome sequence for this species, does not contain any such putative lateral transfer sequences [2,35]. Hence, we are confident that the chloroplast-encoded genes are all encoded in the same chromosome and lack duplicate divergent copies [2] such as those detected by PCR and interpreted as lateral acquisitions [35].

Indeed, the various mitochondrial and nuclear-encoded sequences obtained by single-gene PCR from which a basal position for *Amborella* was once inferred [5] must now be seen in a new light. The interpretations of Bergthorsson *et al.* [35] now raise doubts as to whether any of the previously reported *Amborella* sequences [5] are representative for its true phylogenetic position because they might be lateral acquisitions [35] or otherwise anomalous. With a complete, contiguous chloroplast genome sequence there is no doubt as to which sequences belong to which chromosome and species – a compelling argument in favor of sequencing chloroplast genomes. The conclusion of the *Amborella* debate can be expected when good taxon sampling exists for complete chloroplast genome sequences among angiosperms. Yet, it is reassuring to know that the *Amborella* chloroplast genome is free of PCR-amplifiable duplicate sequences that, for whatever reason, produce phylogenies in which *Amborella* branches all over the green plant tree [35].

Large datasets and 'support'

Soltis *et al.* [1] point out that large datasets can yield high bootstrap values for an erroneous branch or tree but neglect to mention that small datasets can too. Bootstrap proportions do not indicate the confidence that might be placed in a particular branch; they indicate the probability that one would recover the given branch under the specified model given long sequences with the same distribution of site patterns [12]. Matthew Phillips *et al.* [17] made this point succinctly 'Bootstrap support of 100% is not enough; the tree must also be correct. If there are systematic biases, even phylogenetic analysis of complete genomes can be misled by inconsistency'. In datasets encompassing only one or a few genes, sampling errors can lead to high bootstrap values for incorrect branches [12], erroneous support by the measure of bootstraps is by no means unique to large datasets. The use of many different approaches to determine plant phylogeny, including organelle genome data for crucial species such as *Amborella* [2] and *Nymphaea* [3], would seem to be something to welcome, not against which to caution.

Since the first investigations of chloroplast whole genome phylogeny [36] it has been observed that trees appearing to be fully resolved by the measure of bootstrap support are not necessarily correct [16] and that gene sampling is important because different genes analyzed individually can yield different results for the same taxa, often with high bootstrap values [36–38]. Other groups analyzing chloroplast genomes are now observing results that conflict with 100% bootstrap values depending upon the method and taxa employed [11], which underscores the problems that traditional approaches will face as large datasets are assembled.

Conclusion

When we investigate a single gene, we sample a segment of the genome. Comparing more genes reduces the sampling error inherent in just one or a few genes; comparing complete genomes uncovers all the site patterns that are available for comparison. It is well established that genome-scale datasets often contain enough site patterns to convince researchers using one or other model of phylogeny inference that a particular branch is unambiguously resolved. The logical consequence of that realization would seem to be not to advise against the use of genome-scale datasets in evolutionary studies but instead to investigate the properties of the data, the models and the programs that are producing the observed conflicts. If the amount of data were unimportant, we could sequence just one nucleotide from all plants. The more data we have, the less susceptible we are to sampling artefacts inherent to the study of one or a few genes. Conversely, one genome from one plant alone cannot produce a tree.

Genome sequence data can alleviate site-sampling limitations for many levels of phylogenetic resolution. But in doing so, it uncovers the inadequacy, or at least the inconsistency, of many currently employed phylogenetic methods – inadequacies that tend to escape notice with smaller datasets. We consider getting to the root of these issues, dealing with compositional biases and dealing with

covarions to be higher priorities than cautioning against the pursuit of genome sequencing. Traditional approaches to plant phylogeny can generate hypotheses based upon a few genes and many taxa concerning which lineages might be crucial. But those hypotheses cannot be tested effectively without independent data and analyses. Chloroplast genomes from crucial lineages provide needed independent data, the phylogenetic analysis of which entails more issues than taxon sampling alone.

Acknowledgements

We sincerely thank Mikael Thollesson for modifying LDDist to accommodate Dayhoff classes, Peter J. Lockhart for discussions, and the DFG for financial support.

References

1 Soltis, D.E. *et al*. (2004) Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. *Trends Plant Sci.* 9, 477–483
2 Goremykin, V.V. *et al*. (2003) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Mol. Biol. Evol.* 20, 1499–1505
3 Goremykin, V.V. *et al*. (2004) The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. *Mol. Biol. Evol.* 21, 1445–1454
4 Mathews, S. and Donoghue, M.J. (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* 286, 947–950
5 Qiu, Y.L. *et al*. (1999) The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* 402, 404–407
6 Parkinson, C.L. *et al*. (1999) Multigene analyses identify the three earliest lineages of extant flowering plants. *Curr. Biol.* 9, 1485–1488
7 Barkman, T.J. *et al*. (2000) Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. *Proc. Natl. Acad. Sci. U. S. A.* 97, 13166–13171
8 Graham, S.W. and Olmstead, R.G. (2000) Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *Am. J. Bot.* 87, 1712–1730
9 Rosenberg, M.S. and Kumar, S. (2001) Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc. Natl. Acad. Sci. U. S. A.* 98, 10751–10756
10 Lin, Y-H. *et al*. (2002) Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. *Mol. Biol. Evol.* 19, 2060–2070
11 Stefanović, S. *et al*. (2004) Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evol. Biol.* 4, 35
12 Felsenstein, J. (2004) *Inferring Phylogenies*, Sinauer
13 Semple, C. and Steel, M.A. (2003) *Phylogenetics*, Oxford University Press
14 Driskell, A.C. *et al*. (2004) Prospects for building the tree of life from large sequence databases. *Science* 306, 1172–1174
15 Lockhart, P.J. *et al*. (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* 11, 605–612
16 Lockhart, P.J. *et al*. (1999) Spectral analysis, systematic bias, and the evolution of chloroplasts. *Mol. Biol. Evol.* 16, 573–576
17 Phillips, M.J. *et al*. (2004) Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* 21, 1455–1458
18 Ho, S.Y. and Jermiin, L. (2004) Tracing the decay of the historical signal in biological sequence data. *Syst. Biol.* 53, 623–637
19 Lake, J.A. (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proc. Natl. Acad. Sci. U. S. A.* 91, 1455–1459
20 Bryant, D. and Moulton, V. (2004) Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* 21, 255–265
21 Thollesson, M. (2004) LDDist: a Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. *Bioinformatics* 20, 416–418
22 Delsuc, F. *et al*. (2003) Comment on "Hexapod origins: monophyletic or paraphyletic?". *Science* 301, 1482
23 Hrdy, I. *et al*. (2004) *Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* 432, 618–622
24 Naylor, G.J.P. and Brown, W.M. (1997) Structural biology and phylogenetic estimation. *Nature* 388, 527–528
25 Steel, M.A. *et al*. (2000) Invariable site models and their use in phylogeny reconstruction. *Syst. Biol.* 49, 225–232
26 Galtier, N. and Gouy, M. (1998) Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* 15, 871–879
27 Fitch, W.F. and Markowitz, E. (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
28 Penny, D. *et al*. (2001) Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J. Mol. Evol.* 53, 711–723
29 Lockhart, P.J. *et al*. (1998) A covariotide model explains the evolution of oxygenic photosynthesis. *Mol. Biol. Evol.* 15, 1183–1188
30 Lockhart, P.J. *et al*. (2000) How molecules evolve in eubacteria. *Mol. Biol. Evol.* 17, 835–838
31 Philippe, H. *et al*. (2003) Heterotachy and functional shift in protein evolution. *IUBMB Life* 55, 257–265
32 Inagaki, Y. *et al*. (2004) Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1 alpha phylogenies. *Mol. Biol. Evol.* 21, 1340–1349
33 Kolaczkowski, B. and Thornton, J.W. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431, 980–984
34 Ané, C. *et al.* Covarion structure in plastid genome evolution: a new statistical test. *Mol. Biol. Evol.* (in press)
35 Bergthorsson, U. *et al*. (2004) Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm *Amborella*. *Proc. Natl. Acad. Sci. U. S. A.* 101, 17747–17752
36 Goremykin, V. *et al*. (1997) Evolutionary analysis of 58 proteins encoded in six completely sequenced chloroplast genomes: revised molecular estimates of two seed plant divergence times. *Pl. Syst. Evol.* 206, 337–351
37 Martin, W. *et al*. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393, 162–165
38 Vogl, C. *et al*. (2003) Probabilistic analysis indicates discordant gene trees in chloroplast evolution. *J. Mol. Evol.* 56, 330–340
39 Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425
40 Huson, D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14, 68–73
41 Qiu, Y.L. *et al*. (2001) Was the ANITA rooting of the angiosperm phylogeny affected by long-branch attraction? *Amborella*, Nymphaeales, Illiciales, Trimeniaceae, and *Austrobaileya*. *Mol. Biol. Evol.* 18, 1745–1753