The Position of Gnetales among Seed Plants: Overcoming Pitfalls of Chloroplast Phylogenomics

Bojian Zhong,*†, 1 Takahiro Yonezawa, 1 Yang Zhong, 1,2 and Masami Hasegawa* ,1,3

1 School of Life Sciences, Fudan University, Shanghai, China
2 Institute of Biodiversity Science and Geobiology, Tibet University, Lhasa, China
3 Institute of Statistical Mathematics, Tokyo, Japan

†Present address: The Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, 4442, New Zealand.

*Corresponding author: E-mail: bzhong@gmail.com; hasegawa@fudan.edu.cn.
Associate editor: Naruya Saitou

Abstract

The phylogenetic position of Gnetales is one of the most contentious issues in the seed plant systematics. To elucidate the Gnetales position, an improved amino acid substitution matrix was estimated based on 64 chloroplast (cp) genomes and was applied to cp genome data including all three lineages of Gnetales in maximum likelihood analyses of proteins. Although the initial analysis strongly supported the sister relation of Gnetales with Cryptomeria (Cupressophyta or non-Pinaceae conifers) (the “Gnecup” hypothesis), the support seems to be caused by a long-branch attraction (LBA) artifact. Indeed, by removing fastest evolving proteins that are most likely associated with the LBA, the support drastically declined. Furthermore, another analysis of partial genome data with dense taxon sampling of conifers showed that, in psbC, rpl2, and rps7 proteins, there are many parallel amino acid substitutions between the lineages leading to Gnetales and to Cryptomeria, and by further excluding these three genes, the sister relation of Gnetales with Pinaceae (the “Gnepine” hypothesis) became supported. Overall, our analyses indicate that the LBA and parallel substitutions cause a seriously biased inference of phylogenetic position of Gnetales with the cp genome data.

Key words: Gnetales, long-branch attraction, maximum likelihood, model of amino acid substitutions, parallel evolution.

Introduction

The phylogenetic position of Gnetales, a small group of gymnosperms comprised three genera (Ephedra, Gnetum, and Welwitschia), is one of the most controversial issues for the seed plant phylogeny (Chaw et al. 2000; Donoghue and Doyle 2000; Burleigh and Mathews 2004; Mathews 2009). From the early morphological cladistic analyses, Gnetales was considered to be sister to angiosperms (the “Anthophyte” hypothesis) (Doyle and Donoghue 1986; Rothwell and Serbet 1994). On the other hand, most molecular studies do not support the “Anthophyte” hypothesis but do not reach the agreement regarding the position of Gnetales (reviewed in Burleigh and Mathews 2004). Currently, only the three hypotheses receive some support in molecular analyses: 1) Gnetales be placed as sister to conifers as a whole (the “Gnetifer” hypothesis) (e.g., Chaw et al. 1997); 2) within conifers, and sister to Pinaceae (the “Gnepine” hypothesis) (e.g., Bowe et al. 2000; Chaw et al. 2000; Hajibabaei et al. 2006; Wu et al. 2007); and 3) within conifers, but sister to Cupressophyta (non-Pinaceae conifers) (the “Gnecup” hypothesis) (e.g., Nickrent et al. 2000; Doyle 2006; Chumley et al. 2008).

The previous analyses on the position of Gnetales with the multiple genes, however, were problematic and remained to be explored from several aspects as follows: 1) Chloroplast (cp) genomes of Gnetales and non-Pinaceae conifers were poorly sampled with respect to taxa. The genome-scale plastid data sets used previously for this issue included only one species of Gnetales, and no non-Pinaceae conifers was included (Cycas, Ginkgo, Pinus, and Gnetum in Wu et al. 2007; Cycas, Ginkgo, Pinus, and Welwitschia in McCoy et al. 2008); 2) Long-branch attraction (LBA) artifact (Felsenstein 1978; Hendy and Penny 1989; Lockhart and Steel 2005) might be operating. There are several lines of evidence of a long branch leading to Gnetales (Rai et al. 2003; Hajibabaei et al. 2006) that could potentially cause the LBA. The two widely used strategies to alleviate the LBA artifact are to remove most rapidly evolving sequences or sites (Philippe et al. 2005; Hajibabaei et al. 2006; Rai et al. 2008) and to add more taxa (Graybeal 1998; Hillis 1998; Hedtke et al. 2006; Graham and Iles 2009); and 3) Parallel substitutions can potentially cause a problem in inferring the phylogeny. Several cases of parallel substitutions, which referred to the independent evolution of similar traits, starting from the same ancestral character state, have been identified in several studies (Zhang 2006; Rokas and Carroll 2008). Such a phenomenon and convergent molecular evolution could also mislead phylogenetic inference (Christin et al. 2007; Rogozin et al. 2008; Castoe et al. 2009; Li et al. 2010; Liu et al. 2010).

With the aim of addressing these problems, we first analyzed protein-encoding genes in the currently available cp genome data set, which consists of all major groups of seed
plants, including all three genera of Gnetales (McCoy et al. 2008; Wu et al. 2009) and one representative of non-Pinaceae conifers (Hirao et al. 2008), and assessed whether the LBA artifact influences the phylogenetic placement of Gnetales. Next, we carried out several approaches to alleviate the potential biases such as the LBA and the parallel substitutions using the multiple cp genes as well as nuclear genes with more taxa from non-Pinaceae conifers.

Currently, cp genome data are frequently used in evolutionary studies of plants (e.g., Martin et al. 1998; Jansen et al. 2007; Moore et al. 2007; Zhong et al. 2009). If one is interested in resolving orders of deep branchings for the seed plants phylogeny by using cp protein-coding genes, analyses of amino acid sequences are preferable because synonymous substitutions are already saturated and accordingly should have almost no phylogenetic information, and amino acid substitutions are easier to be modeled with an empirical substitution matrix than nucleotide substitutions. For this reason, our analyses in this work are confined to the amino acid sequences of proteins. Although several empirical amino acid substitution models specific to animals are available, only the cpREV model (Adachi et al. 2000) has been reported for plant cp proteins. With the development of the sequencing technique and the rapid accumulation of cp genome data, it seems timely to improve the cpREV model based on an empirical substitution matrix rather than the original cpREV model by using PAML 4 (Yang 2007).

**Materials and Methods**

**Construction of a New Amino Acid Substitution Model**

We retrieved 77 cp protein-encoding genes from 64 taxa, consisting of most of the major angiosperm lineages and three gymnosperms (Cycas, Ginkgo, and Pinus) (Jansen et al. 2007), and the tree topology in Jansen et al. (2007) was assumed. The sites that had ambiguous alignments were excluded, resulting in an alignment of 20199 amino acid positions. Stationary frequencies of amino acids were estimated from the data, and the original cpREV model (Adachi et al. 2000) was applied for the starting parameter estimates. An empirical model of amino acid substitutions was constructed by estimating relative substitution rates between amino acids under the general time-reversible model by using PAML 4 (Yang 2007).

**Data Sets for Phylogenetic Analyses**

The following four data sets were used for the phylogenetic analyses.

1) Data set 1: 56 cp genes from 13 taxa

   This data set comprises the 56 cp protein-encoding genes from all major groups of seed plants (11 taxa) including all three genera of Gnetales, one representative of non-Pinaceae conifers (Cryptomeria japonica) (only data of this species is available from non-Pinaceae conifers) and two representatives of Pinaceae conifers (Keteleeria davidianna and Pinus thunbergii), with Physcomitrella patens and Marchantia polymorpha as outgroups. The details of taxon names and GenBank accession numbers are listed in supplementary table S1, and the 56 gene names are listed in supplementary table S3 (Supplementary Material online). The data set used in estimating the new cp amino acid substitution matrix is mostly independent from Data set 1 and does not contain the phylogenetic issue going to be discussed on the position of Gnetales. Therefore, Data set 1 should be a suitable example to evaluate the performance of the new substitution model.

2) Data set 2a: 14 cp genes from 17 taxa (dense taxon sampling)

   This data set includes the 14 cp protein-encoding genes collected in Graham and Iles (2009) excluding psbN, ndhB, and ndhF because of ambiguous alignments, with additional four taxa (Torreya californica, Saxegothaea conspicua Lindl, Araucaria cunninghamii, and Cedrus deodara) to the 13 taxa in Data set 1.

3) Data set 2b: 14 cp genes from the 13 taxa (sparse taxon sampling)

   This data set is a subset of Data set 2a, with the same taxa as Data set 1.

4) Data set 3: 3 nuclear genes from 12 taxa

   This data set consists of three nuclear protein-encoding genes from the 12 taxa used in Hajibabaei et al. (2006). All the gaps and ambiguous alignments were discarded in all data sets.

**Phylogenetic Inference**

Heuristic tree search with maximum likelihood (ML) phylogenetic analyses were performed by using RAxML 7.0.4 (Stamatakis 2006; Stamatakis et al. 2008) with the cpREV matrix (Adachi et al. 2000) of amino acid substitutions and a discrete gamma distribution with four rate categories. The node support was evaluated based on 100 bootstrap replications. More detailed ML analyses for a limited number of tree topologies were conducted by using CodeML of the PAML 4 package (Yang 2007). In the separate method (Kishino and Hasegawa 1989; Yang 1996), branch lengths and alpha parameters of the gamma distribution could be estimated independently for each protein, whereas in the concatenate method, the concatenated sequences were regarded as homogeneous. To examine whether the performance of phylogenetic inference changed by the improvement of the amino acid substitution model, we conducted ML analyses on the three alternative positions of Gnetales. For Data set 1, Data set 2a, and Data set 2b, the log-likelihood score of each topology was calculated with CodeML of the PAML by using the Dayhoff (Dayhoff et al. 1978), JTT (Jones et al. 1992), cpREV (Adachi et al. 2000), and the new cp substitution matrices with the discrete gamma model for site heterogeneity.

**Removal of Fast-Evolving Proteins**

To reduce the impact of systematic errors (Philippe et al. 2005; Nishihara et al. 2007), we detected and selectively discarded the fast-evolving proteins. For Data set 1 and 2a, the total branch lengths, base frequencies, and alpha parameters of the gamma distribution were estimated...
independently for each protein by using the CodeML program in PAML 4 (Yang 2007) with our new amino acid substitution model. The proteins with long total branch lengths (larger than 2.5 substitutions per amino acid site) were excluded. The total support for a particular tree was evaluated by summing up the log-likelihood scores of each protein, and the total log-likelihood scores of the 3 candidate topologies were compared.

**Reconstruction of Ancestral Character States**

The ancestral site states were inferred for each interior nodes by the empirical Bayesian approach (Yang et al. 1995; Koshi and Goldstein 1996) based on densely sampled Data set 2a using the CodeML program in PAML 4 (Yang 2007) with the new amino acid substitution model. Parallel substitutions were identified by comparing ancestral and extant sequences.

**Results and Discussion**

**The New Amino Acid Substitution Model**

To improve the fitness of the cpREV model, we estimated a new cp amino acid substitution matrix using 77 protein-encoding genes from the 64 sequenced plastid genomes (Jansen et al. 2007), which is by far the most extensive data matrix applied to this issue. The new model is called the cpREV64 model, and Adachi et al. (2000) cpREV model will hereafter be referred to the cpREV10 model associated to the number of the sequences used. The replacement rates of the two cp amino acid substitution models are shown in figure 1, and the estimated amino acid substitution matrix is shown in supplementary table S2 (Supplementary Material online).

**Phylogenetic Analyses and Model Comparison Based on the 56 cp Proteins from the 13 Taxa**

From the ML analysis with a partitioned approach using Data set 1, the monophyly of Gnetales and Cryptomeria (non-Pinaceae conifers) is strongly supported (100% bootstrap support) (fig. 2). However, the branches leading to Gnetales and Cryptomeria are both significantly longer than those of other branches (see fig. 2 and supplementary fig. S1, Supplementary Material online), so we suspected that the Gnetales placement might reflect a LBA artifact (Felsenstein 1978; Hendy and Penny 1989).

In order to obtain a more reliable placement of Gnetales, and to evaluate the utility of various amino acid models, we applied ML analyses to compare the efficiency of amino acid models (including the cpREV64 model) with the discrete $I$ distribution for site heterogeneity. Table 1 shows that the Gnetales/Cryptomeria clade (Tree 1: the “Gnecup” hypothesis) was strongly supported no matter which amino acid substitution model was used and that alternative trees of the Gnetales/Pinaceae clade (Tree 2: the “Gnepine” hypothesis) and the Gnetales/conifer clade (Tree 3: the “Gnetifer” hypothesis) were strongly rejected. However, there is a trend of decreasing the support of Tree 1 as the fitting of the data increases. For instance, the $P$ value of the AU test (Shimodaira 2002) for Tree 2 is highest with the cpREV64+$I$ model ($3 \times 10^{-4}$) among those of alternative matrices, indicating that Tree 2 was relatively preferred with the improvement of the matrix. Indeed, the cpREV64+$I$ model is the best model among the alternatives used in this analysis because the model gave the highest likelihood scores. It should be noted that the data set we analyzed for the phylogenetic placement of Gnetales is almost independent from the data set used in estimating the cpREV64 matrix. It may indicate that the cpREV64 matrix could relatively alleviate the LBA artifact resulting in reducing the support of Tree 1.

The phylogenomics approach using a genome-scale data is thought to be useful in estimating robust phylogenies (Philippe and Telford 2006). Several empirical studies have confirmed that the use of large amount of data could reduce the impact of stochastic error and could overcome incongruence (e.g., Baptiste et al. 2002; Philippe et al.
However, conflicting results have also been reported, the support of erroneous trees can be enhanced by simply combining sequences (Gadagkar et al. 2005; Nishihara et al. 2007). Furthermore, the impact of systematic errors, such as the LBA artifact and the site substitution rate change over time, are amplified in the genome-scale approach. Our phylogenetic analyses of Gnetales based on the genome-scale cp data set also suggest that the LBA artifact could have biased the phylogenetic inference.

Impact of Removing Fast-Evolving Proteins for Phylogenetic Inference

It has become evident that gathering a large amount of data is not sufficient enough to obtain a reliable tree and we must try to alleviate causes that lead us to biased tree estimation. One of the well-known factors to cause biased tree estimation is the LBA artifact, and fast evolutionary rates are often associated with such a misleading effect. In order to check the possibility of the LBA artifact more in detail, we removed the 18 fastest evolving proteins (of 56 proteins) for which the total branch lengths were larger than 2.5 substitutions per site. Detailed branch lengths for each protein are shown in supplementary table S3 (Supplementary Material online). Analyses were carried out with ML and with both the concatenate and separate methods (table 2).

When all the proteins in Data set 1 were included, Tree 1 was supported with the highest log-likelihood scores. In contrast, Tree 2 could not be rejected (P = 0.214 by AU test) using concatenate method after removing the 18 fastest proteins (Data Set 1b). Indeed, Tree 2 became to be preferred with the highest log-likelihood scores (−48799.82) when the separate method was applied, even though Tree 1 could not be excluded. Removing the rapidly evolving sites or sequences has been suggested to improve the accuracy of phylogenetic inference (e.g., Rydin et al. 2002; Philippe et al. 2005; Rodriguez-Ezpeleta et al. 2007). Our results mentioned above may suggest that the fast-evolving proteins have caused biased estimation supporting the Gnetales/Cryptomeria cluster due to the LBA artifact, and the deletion of the rapidly evolving proteins has a notable effect on the phylogenetic placement of Gnetales.

Impact of Dense Taxon Sampling for Phylogenetic Placement of Gnetales

Dense taxon sampling has been proved to be another approach to break up long branches and to alleviate the impact of the LBA (e.g., Graybeal 1998; Hillis 1998; Hedtke 2004). However, conflicting results have also been reported, the support of erroneous trees can be enhanced by simply combining sequences (Gadagkar et al. 2005; Nishihara et al. 2007). Furthermore, the impact of systematic errors, such as the LBA artifact and the site substitution rate change over time, are amplified in the genome-scale approach. Our phylogenetic analyses of Gnetales based on the genome-scale cp data set also suggest that the LBA artifact could have biased the phylogenetic inference.

Impact of Removing Fast-Evolving Proteins for Phylogenetic Inference

It has become evident that gathering a large amount of data is not sufficient enough to obtain a reliable tree and we must try to alleviate causes that lead us to biased tree estimation. One of the well-known factors to cause biased tree estimation is the LBA artifact, and fast evolutionary rates are often associated with such a misleading effect. In order to check the possibility of the LBA artifact more in detail, we removed the 18 fastest evolving proteins (of 56 proteins) for which the total branch lengths were larger than 2.5 substitutions per site. Detailed branch lengths for each protein are shown in supplementary table S3 (Supplementary Material online). Analyses were carried out with ML and with both the concatenate and separate methods (table 2).

When all the proteins in Data set 1 were included, Tree 1 was supported with the highest log-likelihood scores. In contrast, Tree 2 could not be rejected (P = 0.214 by AU test) using concatenate method after removing the 18 fastest proteins (Data Set 1b). Indeed, Tree 2 became to be preferred with the highest log-likelihood scores (−48799.82) when the separate method was applied, even though Tree 1 could not be excluded. Removing the rapidly evolving sites or sequences has been suggested to improve the accuracy of phylogenetic inference (e.g., Rydin et al. 2002; Philippe et al. 2005; Rodriguez-Ezpeleta et al. 2007). Our results mentioned above may suggest that the fast-evolving proteins have caused biased estimation supporting the Gnetales/Cryptomeria cluster due to the LBA artifact, and the deletion of the rapidly evolving proteins has a notable effect on the phylogenetic placement of Gnetales.

Impact of Dense Taxon Sampling for Phylogenetic Placement of Gnetales

Dense taxon sampling has been proved to be another approach to break up long branches and to alleviate the impact of the LBA (e.g., Graybeal 1998; Hillis 1998; Hedtke

### Table 1. Model Comparisons of the Three Alternative Placements of Gnetales Based on Data Set 1.

<table>
<thead>
<tr>
<th>Models</th>
<th>Tree 1: Gneucup</th>
<th>Tree 2: Gnepine</th>
<th>Tree 3: Gnetifer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dayhoff + Γ</td>
<td>−127897.02</td>
<td>−136.50 (&lt;10&lt;sup&gt;−6&lt;/sup&gt;)</td>
<td>−130.40 (3 × 10&lt;sup&gt;−5&lt;/sup&gt;)</td>
</tr>
<tr>
<td>JTT + Γ</td>
<td>−124331.23</td>
<td>−119.19 (&lt;10&lt;sup&gt;−6&lt;/sup&gt;)</td>
<td>−119.62 (&lt;10&lt;sup&gt;−6&lt;/sup&gt;)</td>
</tr>
<tr>
<td>cpREV10 + Γ</td>
<td>−124646.39</td>
<td>−119.42 (&lt;10&lt;sup&gt;−6&lt;/sup&gt;)</td>
<td>−115.99 (&lt;10&lt;sup&gt;−6&lt;/sup&gt;)</td>
</tr>
<tr>
<td>cpREV64 + Γ</td>
<td>−122991.77</td>
<td>−103.82 (3 × 10&lt;sup&gt;−4&lt;/sup&gt;)</td>
<td>−104.30 (&lt;10&lt;sup&gt;−6&lt;/sup&gt;)</td>
</tr>
</tbody>
</table>

Note.—The log-likelihood of the ML tree (in angled brackets) and the differences in log-likelihood of alternative trees from that of the ML tree are shown by the concatenate method. The P values of the AU test are given in parentheses. The following three tree topologies are examined.

Tree 1: “Gneucup” hypothesis: Gnetales sister to non-Pinaceae conifers.
Tree 2: “Gnepine” hypothesis: Gnetales sister to Pinaceae.
Tree 3: “Gnetifer” hypothesis: Gnetales sister to conifers.
Gnetales contain only three extant genera and others are all extinct, so it is impossible to add more genera for Gnetales. In this study, we increased the number of samples from Cupressophyta (non-Pinaceae conifers) and Pinaceae groups, with three additional taxa (T. californica, S. conspicua Lindl, A. cunninghamii) from Cupressophyta, as well as one taxon (C. deodara) from Pinaceae. Although the whole cp genome data have not been sequenced from these additional taxa, a number of genes are publicly available (Graham and Iles 2009). To evaluate the performance of dense taxon sampling for the phylogenetic accuracy, the comparison of the three alternative Gnetales positions based on the dense sampling (Data sets 2a) and the sparse sampling (Data sets 2b) were conducted (table 3). When the sparsely sampled data were applied, Tree 1 was strongly supported as well as the strong rejection of Tree 2 and Tree 3. In contrast, although Tree 1 constantly had the highest log-likelihood scores, Tree 2 and Tree 3 could not be rejected (P value: 0.471 and 0.304, respectively, by AU test) based on the dense sampling, and the log-likelihood differences among the three alternative trees were very minor. In addition, the long branch to Cryptomeria in the sparse sampling was cut into short branches in the dense sampling which may have contributed to reduce the bias of tree inference (supplementary fig. S2, Supplementary Material online).

### Table 2. Comparison of the Log-Likelihood for the Three Alternative Trees with Concatenate and Separate Methods from the 13 Taxa.

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>Concatenate Method</th>
<th>Separate Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tree 1: Gnceup</td>
<td>Tree 2: Gnepine</td>
</tr>
<tr>
<td>Data set 1a: all proteins</td>
<td>$&lt;-122991.77$</td>
<td>$-103.82$ ($3 \times 10^{-4}$)</td>
</tr>
<tr>
<td>Data set 1b: fast-evolving proteins excluded</td>
<td>$&lt;-50652.37$</td>
<td>$-12.21$ (0.214)</td>
</tr>
<tr>
<td>Data set 1c: psbC, rpl2 and rps7 further excluded</td>
<td>$-10.84$ (0.225)</td>
<td>$&lt;44385.20$</td>
</tr>
</tbody>
</table>

**Note:** The log-likelihood of the ML tree (in angled brackets) and the differences in log-likelihood of alternative trees from that of the ML tree are shown. The P values of the AU test are given in parentheses and the Akaike information criterion (AIC) values are also shown for Data set 1c.

### Table 3. Comparison of the Log-Likelihood for the Three Alternative Trees Based on Dense Sampling and Sparse Sampling Using the cpREV64 Model.

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>Tree 1: Gnceup</th>
<th>Tree 2: Gnepine</th>
<th>Tree 3: Gnetifer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set 2b (sparse sampling)</td>
<td>$&lt;-15972.28$</td>
<td>$-27.90$ (0.003)</td>
<td>$-28.07$ (0.003)</td>
</tr>
<tr>
<td>Data set 2a (dense sampling)</td>
<td>$&lt;-17193.18$</td>
<td>$-1.19$ (0.471)</td>
<td>$-2.67$ (0.304)</td>
</tr>
</tbody>
</table>

**Note:** The log-likelihood of the ML tree (in angled brackets) and the differences in log-likelihood of alternative trees from that of the ML tree are shown. The P values of the AU test are given in parentheses.

Phylogenetic Analysis Based on Nuclear Proteins

Cp genomic data have been used in most recent molecular studies to resolve the position of Gnetales (Chaw et al. 1997, 2000; Wu et al. 2007; McCoy et al. 2008), but an additional analysis of nuclear data should be helpful in clarifying this problem (Hajibabaei et al. 2006). To compare the phylogenetic position of Gnetales inferred from the nuclear genes with that from the cp genes, we conducted ML analyses based on the nuclear proteins (Data set 3). Remarkably, none of the four models supported Tree 1, and the JTT+$I^+$ turned to be the best model (the highest log-likelihood) for the nuclear data, and preferred Tree 2 (although marginally, table 4). Therefore, combining our cp analyses with this independent evidence from the nuclear data, the “Gnceup” hypothesis (Tree 1) seems unreliable, and the “Gnepine” hypothesis (Tree 2) appears to be preferred.

Parallel Amino Acid Replacements and Biased Inference of the Phylogenetic Position of Gnetales

When the total branch lengths of each protein in Data set 1 were estimated, most of the proteins had a longer branch to Cryptomeria than to Pinaceae, but only psbC protein held an almost zero branch length of Cryptomeria, whereas the branch leading to Pinaceae had a moderate length (data not shown).

To elucidate the abnormal behavior of psbC protein, we estimated the ancestral site state of psbC and other proteins based on densely sampled Data set 2a and identified as many as 13 parallel amino acid substitutions that occurred between the branch leading to Cryptomeria and the common ancestral branch of Gnetales in psbC, rpl2, and rps7 proteins, whereas only one parallel amino acid substitution was found in atpB protein between the branch leading to Pinus and the ancestral Gnetales branch (fig. 3 and details in table 5). We suspected that the parallel substitutions may have strongly influenced the placement of Gnetales with Data set 1 (only Cryptomeria from Cupressophyta was included in Data set 1), the Gnetales/Cryptomeria grouping is likely to be preferred due to the excess of parallel amino acid changes between the branches leading to Gnetales and Cryptomeria in the three proteins (psbC, rpl2, and rps7). To reduce the bias from the parallel substitutions, we further screened out these three proteins from the data set that excluded the fast-evolving proteins and carried out ML analyses by concatenate as well as separate methods (Data Set 1c in table 2). Intriguingly, Tree 2 was favored over the alternative hypotheses with both methods, and the separate method was superior to the concatenate method based on the Akaike information criterion (Akaike 1973). Furthermore, the ML tree inference based on the Data Set 1c supported the monophyly of Gnetales and Pinaceae with a moderate bootstrap probability.
Table 4. Comparison of the Log-Likelihood for the Three Alternative Trees Based on Nuclear Data (Data set 3).

<table>
<thead>
<tr>
<th>Tree 1: Gneup</th>
<th>Tree 2: Gnepine</th>
<th>Tree 3: Gnetifer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dayhoff + I’</td>
<td>-3.04 (0.039)</td>
<td>-5869.80</td>
</tr>
<tr>
<td>JTT + I’</td>
<td>-3.20 (0.044)</td>
<td>-5832.82</td>
</tr>
<tr>
<td>cpREV10 + I’</td>
<td>-3.03 (0.044)</td>
<td>-5893.95</td>
</tr>
<tr>
<td>cpREV64 + I’</td>
<td>-3.81 (0.034)</td>
<td>-0.08 (0.528)</td>
</tr>
</tbody>
</table>

NOTE.—The log-likelihood of the ML tree (in angled brackets) and the differences in log-likelihood of alternative trees from that of the ML tree are shown by the separate method. The $P$ values of the AU test are given in parentheses.

Table 5. List of Parallel Amino Acid Substitutions.

<table>
<thead>
<tr>
<th>Ancestral Branch to</th>
<th>Proteins</th>
<th>Parallel Substitutions with Ancestral Gnetales Branch</th>
<th>Site Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cryptomeria</td>
<td>psbC</td>
<td>10 A → T</td>
<td>97.2; *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>196 V → T</td>
<td>99.2; *</td>
</tr>
<tr>
<td></td>
<td>rpl2</td>
<td>62 V → I</td>
<td>97.8; *</td>
</tr>
<tr>
<td></td>
<td>rps7</td>
<td>7 A → E</td>
<td>99.0; *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>87 R → K</td>
<td>91.8; *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>101 G → E</td>
<td>99.1; *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>138 R → K</td>
<td>95.0; *</td>
</tr>
<tr>
<td>Torreya/Cryptomeria</td>
<td>PsbC</td>
<td>173 V → I</td>
<td>99.7; 98.8</td>
</tr>
<tr>
<td></td>
<td>rpl2</td>
<td>85 M → I</td>
<td>99.6; 99.0</td>
</tr>
<tr>
<td></td>
<td>rps7</td>
<td>156 G → S</td>
<td>97.8; 99.8</td>
</tr>
<tr>
<td>Pinus</td>
<td>atpB</td>
<td>65 S → P</td>
<td>99.6; *</td>
</tr>
<tr>
<td>Keteleeria</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Pinus/Keteleeria</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>

NOTE.—Only branches leading to the species used in the cp genome analysis are shown. A list for other branches is given in supplemental table S4 (Supplementary Material online). Branch labels refer to those in figure 3. The probabilities of ancestral nodes are shown in % (* refers to a terminal). The numbering of amino acid sites experiencing parallel substitutions are those of Pinus thunbergii (Wakasugi et al.1994).

Impact of Amino Acid Compositional Bias

Amino acid compositions of proteins differ among different taxa, and the difference may cause a biased estimation of the phylogeny (Lockhart et al. 1994; Foster 2004; Phillips et al. 2004). The distances of cp amino acid composition (defined by Adachi and Hasegawa 1996) are shown in table S5. For the fast + parallel genes, the distances of outgroup...
species (Amborella, Cycas, and Ginkgo) to Gnetales are much longer than those to Pinaceae (Pinus and Keteleeria), and the distance to Cupressophyta (Cryptomeria) is intermediate. This trend is held for Data set 1c (exclusion of fast + parallel genes), although in a more moderate way. The extreme difference of amino acid compositions in Gnetales from those in Pinaceae is likely to be one of the reasons why Data set 1a gave a strong support for Tree 1. Furthermore, the retention of the same trend even in a more moderate way after excluding the fast + parallel genes may partially explain why Data set 1c cannot exclude Tree 1 with a high significance.

Impact of Alternative Outgroup Relationship of Cycas and Ginkgo
Cycas and Ginkgo are strongly monophyletic from the ML tree (fig. 2), and therefore we assumed this clade in the subsequent analyses. However, several molecular analyses suggest that Cycas is the basal Gymnosperm lineage (e.g., Bowe et al. 2000; Chaw et al. 2000). To check whether the alternative relationship between Cycas and Ginkgo could affect the position of Gnetales, we further compared the log-likelihood of the three alternative positions of Gnetales with Cycas as the basal Gymnosperm for the three data sets (table S6). The results indicate that “Gnepine” tree is favored by excluding fast-evolving and parallel genes, which are congruent with the analyses assumed the monophyly of Cycas and Ginkgo, and the relative supports of the three alternative trees remain essentially unchanged from those of table 2. So the support of “Gnepine” tree is robust even though alternative relationship between Cycas and Ginkgo is assumed.

Conclusion
Phylogenetic position of Gnetales has been a mysterious issue in the seed plants phylogenetics. With the cp amino acid substitution matrix based on the largest data set for this purpose, we are able to improve the current substitution model to help resolve this controversial problem in this study. We found that the LBA artifact and the parallel changes play significant roles in misleading the phylogenetic placement of Gnetales. In particular, our analyses suggest that parallel amino acid changes have a strong impact in grouping Gnetales with Cryptomeria (the “Gnepine” hypothesis). Here, we have demonstrated that removing fast-evolving genes and increasing taxon sampling can effectively alleviate the LBA artifact, thereby recovering a sister-group position of Gnetales to Pinaceae (the “Gnepine” hypothesis). Furthermore, this hypothesis is increasingly supported by removing the parallel-evolving proteins. Additionally, independent lines of evidence supporting the “Gnepine” hypothesis were recently provided by plastid
structural losses, such as the loss of all ndh genes
(Braukmann et al. 2009) and the loss of rps16 (Wu et al.
2007, 2009) commonly observed between Gnetales and
Pinaceae lineages.

Therefore, although the aberrant phylogenetic signals in
our data sets prevented us to draw a firm conclusion on the
position of Gnetales, the congruence of our analyses makes
us confident in supporting the “Gnepine” hypothesis. Cur-
cently, cp genome data are available from only one species
(C. japonica) among non-Pinaceae conifers. Additional cp
genomic data of the non-Pinaceae conifers is expected to
shed more light on the position of Gnetales.

Supplementary Material
Supplementary tables S1–S6 and figures S1–S4 are available
at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
We would like to thank David Penny for helpful comments.
We also thank the associate editor and two anonymous
reviewers for constructive suggestions on the manuscript.
This work was supported by National Science Foundation
of China (30925004).

References
Adachi J, Hasegawa M. 1996. MOLPHY version 2.3: programs for
molecular phylogenetics based on maximum likelihood. Compt
phylogeny and a model of amino acid substitution for proteins
Akaike H. 1973. Information theory and an extension of the
Second International Symposium on Information Theory.
analysis of 100 genes supports the grouping of three highly
divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba.
based on all three genomic compartments: extant gymnosperms
are monophyletic and Gnetales’ closest relatives are conifers.
Braukmann TW, Kuzmina M, Stefanovic S. 2009. Loss of all plastid
ndh genes in Gnetales and conifers: extent and evolutionary
from seed plants: implications for resolving the seed plant tree of
Castoe TA, de Koning AP, Kim HM, Gu W, Noonan BP, Naylor G,
Jiang ZJ, Parkinson CL, Pollock DD. 2009. Evidence for an ancient
adaptive episode of convergent molecular evolution. Proc Natl
Seed plant phylogeny inferred from all three plant genomes:
monophyly of extant gymnosperms and origin of Gnetales from
phylogeny of extant gymnosperms and seed plant evolution:
C₃ photosynthesis evolved in grasses via parallel adaptive
Gnetalean affinities in seed plant phylogeny with 83 plastid
genes. Botany 2008: Joint Annual Meeting of Canadian Botanical
AssociationAmerican Fern Society, American Society of Plant
Taxonomists, and the Botanical Society of America. Vancouver,
British Columbia (Canada). Available form: http://2008.botany-
Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary
change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and
Donoghue MJ, Doyle JA. 2000. Seed plant phylogeny: demise of the
Doyle JA, Donoghue MJ. 1986. Seed plant phylogeny and the origin of
Felsenstein J. 1978. Cases in which parsimony and compatibility
methods will be positively misleading. Syst Zool. 27:401–410.
53:485–495.
Gadagkar SR, Rosenberg MS, Kumar S. 2005. Inferring species
phylogenies from multiple genes: concatenated sequence tree
Graham SW, Iles WJD. 2009. Different gymnosperm outgroup have
(mostly) congruent signal regarding the root of flowering plant
Graybeal A. 1998. Is it better to add taxa or characters to a difficult
Hajibabaei M, Xia J, Drouin G. 2006. Seed plant phylogeny:
gnetophytes are derived conifers and a sister group to Pinaceae.
Hedtke SM, Townsend TM, Hillis DM. 2006. Resolution of
phylogenetic conflict in large data sets by increased taxon
Hendy M, Penny D. 1989. A framework for the qualitative study of
Hillis DM. 1998. Taxonomic sampling, phylogenetic accuracy, and
nucleotide sequence of the Cryptomeria japonica D. Don.
chloroplast genome and comparative chloroplast genomics:
diversified genomic structure of coniferous species. BMC Plant
Biol. 8:70.
of 81 genes from 64 plastid genomes resolves relationships in
angiosperms and identifies genome-scale evolutionary patterns.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of
mutation data matrices from protein sequences. CABIOS.
8:275–282.
likelihood estimate of the evolutionary tree topologies from
DNA sequence data, and the branching order in hominoidea. J
Koshi JM, Goldstein RA. 1996. Probabilistic reconstruction of
Convergent sequence evolution between echolocating bats and


