GENE SAMPLING VERSUS TAXON SAMPLING: IS AMBORELLA (AMBORELLACEAE) A SISTER GROUP TO ALL OTHER EXTANT ANGIOSPERMS?

Г. В. ДЕГТЯРЕВА, Т. Х. САМИГУЛЛИН, Д. Д. СОКОЛОВ, К. М. ВАЛЬЕХО-РОМАН.
ЯВЛЯЕТСЯ ЛИ AMBORELLA (AMBORELLACEAE) СЕСТРИНСКОЙ ГРУППОЙ ПО ОТНОШЕНИЮ КО ВСЕМ ОСТАЛЬНЫМ СОВРЕМЕННЫМ ПОКРЫТОСЕМЕННЫМ РАСТЕНИЯМ?

© G. V. Degtjareva, Т. H. Samigullin, D. D. Sokoloff, С. M. Valiejo-Roman

The majority of recently published molecular phylogenetic analyses suggest a basal position for Amborella within angiosperms. However, these analyses were based on a relatively limited set of DNA regions. Goremykin et al. (2003) analysed angiosperm phylogeny using sequences of 61 protein-coding genes common to plastomes of land plants. They revealed dicots as sister to monocots and Amborella as sister to Cufyaanthus. We suggest that the unusual tree topology published by Goremykin et al. (2003) reflects low taxon sampling rather than increased gene sampling. We consider that the data inferred from the analyses of 61 chloroplast genes do not refute the idea of a basal position for Amborella. However, we realise that the basal position of Amborella is nothing more than a hypothesis. Although current data on chloroplast genomes, in our opinion, are insufficient to falsify this hypothesis, it may well be falsified by further studies.

Key words: Molecular phylogeny, root of Angiosperms, Amborella, basal monocots.

The problem of relationships among basal groups of Angiosperms is one of most complicated and most disputable problems of plant phylogeny. The use of molecular data allowed the establishment of new hypotheses concerning basal groups of Angiosperms. The analysis of rbcL sequences suggested a basal position for Ceratophyllum, a highly specialized and reduced water plant (Chase et al., 1993). During recent years, Angiosperm phylogeny has been studied using sequences of several different DNA regions. The majority of combined analyses suggest a basal position for Amborella, a shrubby plant endemic to New Caledonia (e. g., Qiu et al., 2000; Soltis et al., 2001; Zanis et al., 2002; APG II, 2003; Hilu et al., 2003; Löhne et al., 2003; see however: Barkman et al., 2000). These analyses allow the recognition of basal Angiosperm grade (ANITA grade), consisting of three clades (1) monotypic family Amborellaceae (2) families Cabombaceae and Nymphaeaceae, and (3) Trimeniaceae, Illiciaceae, and Schisandraceae.

The basal position of Amborella is not in agreement with the traditional viewpoint that the family Amborellaceae represents a basal member of the order Laurales (e. g., Takhta-
After publication of molecular data, some morphological evidence was found in favour of placement of *Amborella* together with other members of the ANITA grade, e.g. from study of gynoecium morphology (Endress, Igersheim, 2000; Endress, 2001). However, new morphological differences of *Amborella* from other members of the ANITA grade were found in female gametophyte structure (see, e.g., Friedman et al., 2003).

At present, only a minor part of plant genome has been used to reconstruct phylogeny of angiosperms. Besides, only a minor part of more than 250 000 angiosperm species (Takhtajan, 1987, 1997) was included in each particular phylogenetic analysis. It is therefore uncertain how realistic is the present day estimation of angiosperm phylogeny inferred from the molecular data set. It is important to test the current hypothesis of angiosperm root by analysis of increased molecular data set. However, it is very difficult to increase simultaneously taxon sampling and gene sampling.

Goremykin et al. (2003) conducted analyses of basal angiosperm relationships based on complete sequences of chloroplast genomes (fig. 1). Using these complete sequences allowed maximum increase of characters involved into analyses. In total, 61 protein-coding genes common to the plastomes of land plants were analysed using neighbour-joining, maximum parsimony and maximum likelihood approaches (Goremykin et al., 2003). In the majority of analyses, *Amborella* was not revealed as the sister group of all other Angiosperms (Goremykin et al., 2003). Rather, *Amborella* clustered with *Calycanthus* (Calycanthaceae). The family Calycanthaceae as well as *Amborellaceae* was traditionally placed in the order Laurales (e.g., Takhtajan, 1987, 1997; Cronquist, 1988; Dahlgren, 1989). Monocots and dicots are highly supported as two sister monophyletic clades in analyses of 61 plastome genes (fig. 1) (Goremykin et al., 2003). This result is in contrast with vast majority of recent molecular analyses. It should be stressed that the mono-

![Fig. 1. Neighbour-Joining tree built from Kimura two-parameter distances derived from analysis of the alignment of the first and the second codon positions from 61 protein-coding genes common to the plastomes of land plants (redrawn from Goremykin et al., 2003).](image-url)

10 angiosperm taxa are analysed. Bootstrap support indices are shown.
Fig. 2. Phylogenetic relationships among basal groups of angiosperms inferred from analysis of 18S rDNA, rbcL, and atpB sequences in 567 angiosperm taxa. Parsimony ratchet is used. Jackknife support indices are shown. Simplified from Soltis et al. (2000).

Phylogeny of dicots (in a cladistic sense!) disagrees also with the traditional viewpoint based on study of morphological data (e.g., Takhtajan, 1966, 1987, 1997; Dahlgren et al., 1985; Cronquist, 1988).

According to Goremykin et al. (2003) their result implies that the phylogeny inferred from studies of full chloroplast genomes is highly incongruent with the phylogeny inferred from analyses of a few genes. However, Goremykin et al. (2003) included only 13 taxa in their analyses (three outgroups and ten angiosperm taxa). In contrast, some recent analyses, which demonstrated the basal position of Amborella, included more than hundred plant species. It might be possible that the unusual tree topology obtained by Goremykin et al. (2003) reflected low taxon sampling rather than increased gene sampling. To test such a hypothesis, we have analysed almost the same set of taxa as Goremykin et al. (2003) using data for only three genes, rbcL, atpB, and 18S rDNA. These genes were analysed previously by Soltis et al. (2000), who used data for a large sample of angiosperm taxa. In the analysis by Soltis et al. (2000), Amborella was revealed as the basalmost angiosperm (fig. 2). This test should allow demonstration of the impact of decreased taxon sampling on tree topology.

The ten ingroup genera analysed by Goremykin et al. (2003) do not represent all important angiosperm groups recognised in traditional classifications. In particular, monocots are represented in these analyses exclusively by grasses (Zea, Triticum, and Oryza). We have tried to test what impact on tree topology different sampling of monocots may have. To do this, we have conducted analyses of three gene data set for 13 genera, among which grass taxa were replaced by three more basal monocot representatives, Japonolirion, Tofieldia, and Lilium. These monocot genera exhibit a number of putative plesiomorphic features of floral morphology (especially when compared to grasses).

Methods

GenBank data were used to construct two matrices each containing 13 genera and sequences of three genes (rbcL, ATP, 18S rDNA). The first matrix included almost the same set of taxa as was analysed by Goremykin et al. (2003). Due to technical reasons, Clar-
<table>
<thead>
<tr>
<th>Genus</th>
<th>18S rDNA GenBank accession number and species</th>
<th>18S rDNA First publication of the sequence</th>
<th>18S rDNA Gen Bank accession number and species</th>
<th>18S rDNA First publication of the sequence</th>
<th>rpsL GenBank accession number and species</th>
<th>rpsL First publication of the sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oryza</td>
<td>X00755 0. sativa L.</td>
<td>Soltis et al. (1997)</td>
<td>X15901 0. sativa L.</td>
<td>Chase et al. (1993)</td>
<td>X15901 0. sativa L.</td>
<td>Chase et al. (1993)</td>
</tr>
</tbody>
</table>
**Table continued**

<table>
<thead>
<tr>
<th>Genus</th>
<th>First publication of the sequence</th>
<th>GenBank accession number and species</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>rbcL</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>atpB</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>18S rDNA</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GenBank</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Kia was used instead of Oenothera as a representative of Onagraceae and Pisum is used instead of Lotus as a representative of Leguminosae. The second matrix differs from the first matrix by monocot sampling. Japonolirion, Tofieldia, and Lilium are included in the second matrix instead of Zea, Triticum, and Oryza. GenBank accession numbers for all sequences used in our study are presented in the Table.

The DNA sequences were aligned initially using CLUSTAL W version 1.75 (Thompson et al., 1994) and corrected manually using the SED editor of the VOSTORG phylogenetic analysis package (Zharkikh et al., 1990). The alignment is available from the authors upon request.

Maximum parsimony analyses were performed using PAUP* version 4.0b8 (Swofford, 2000) and WinClada version 1.00.08 (Nixon, 2002). In the first case (as in study by Goremykin et al., 2003), bootstrap support indices were calculated, while in the second case (as in study by Soltis et al., 2000) jackknife support indices and ratchet algorithm were used.

Distance trees were calculated using the neighbour-joining method (Saitou, Nei, 1987) as it is implemented in PAUP*. Kimura two-parameter distances were calculated since this model was used by Goremykin et al. (2003). 1000 bootstrap resamplings were performed.

Maximum likelihood analyses were performed using TreePuzzle (Schmidt et al., 2000) and MrBayes (Huelsenbeck, Ronquist, 2001) programs. Goremykin et al. (2003) used TreePuzzle software only.

**Results**

**Neighbour-joining analysis** of three gene data set (rbcL, atpB, and 18S rDNA) for almost the same taxon set as used by Goremykin et al. (2003) revealed almost the same tree topology as was published by Goremykin et al. (2003) (fig. 1, 3). Dicots are 100 % supported as a monophyletic group and Amborella is 100 % supported as sister to Calycanthus in both analyses. When Tofieldia, Japonolirion, and Lilium represented monocots, neighbour-joining analysis of the three gene data set also did not reveal a basal position for Amborella (fig. 4).

**Maximum parsimony analyses** of three gene/13 taxon matrices demonstrated no significant differences between the use of PAUP* and WinClada programs. Maximum parsimony analyses of three gene data set for almost the same taxon set as was used by Goremykin et al. (2003) revealed monophyly of dicots (fig. 5). As well as in study by Goremykin et al. (2003), Amborella is not revealed as the basal most angiosperm. However, in contrast to results by Goremykin et al.
Fig. 3. Neighbour-Joining tree built from Kimura two-parameter distances derived from analysis of the alignment of all codon positions from three genes, rbcL, atpB, and 18S rDNA. These three genes were analysed previously by Soltis et al. (2000), who used data for a large sample of angiosperm taxa and concluded that *Amborella* represented a sister group to all other extant Angiosperms. In the analysis with limited taxon sampling we have obtained almost the same tree topology as Goremykin et al. (2003), with *Amborella* sister to *Calycanthus* (Calycanthaceae, Laurales).

(fig. 1), *Amborella* is not clustered with *Calycanthus* (fig. 5). When *Tofieldia*, *Japonolirion* and *Lilium* represented monocots, maximum parsimony analyses of the three gene data set revealed *Amborella* as a basal clade within angiosperms; the rest of angiosperms formed a clade with 98 % jackknife support and 100 % bootstrap support (fig. 6).

**Maximum likelihood analyses using TreePuzzle program.** When grasses represented monocots, *Amborella* is not revealed as the basalmost angiosperm (fig. 7). As in the tree published by Goremykin et al. (2003), *Amborella* is clustered with *Calycanthus* (figs. 1, 7). When *Tofieldia*, *Japonolirion*, and *Lilium* represented monocots, *Amborella*

Fig. 4. Neighbour-Joining tree built from Kimura two-parameter distances derived from analysis of the alignment of all codon positions from three genes, rbcL, atpB, and 18S rDNA. Monocots are represented by three genera with relatively primitive floral structure (*Tofieldia*, *Japonolirion*, *Lilium*).
Fig. 5. Maximum parsimony tree derived from analyses of the alignment of all codon positions from three genes, *rbcL*, *atpB*, and 18S rDNA. Monocots are represented only by grasses in this analysis (as well as in the analyses by Goremykin et al., 2003). *Amborella* is not revealed as the basalmost angiosperm. Numbers above branches are jackknife support values found in analysis with ratchet algorithm (the method used by Soltis et al., 2000). Numbers below branches are bootstrap support values found by PAUP* (the method used by Goremykin et al., 2003). Bootstrap support less than 50% is not shown.

is revealed as the basal most angiosperm (fig. 8), and the whole tree topology is compatible with the tree published by Soltis et al. (2000) (see fig. 2).

**Bayesian analyses.** Maximum likelihood analyses of the three gene data sets using MrBayes program revealed a basal position for *Amborella* (fig. 9). In contrast to the rest of the analyses performed in the present study, the position of *Amborella* does not depend on monocot sampling. The clade containing all analysed angiosperms minus *Amborella* received 100% posterior probability when grasses represented monocots as well as when *Tofieldia*, *Japonolirion*, and *Lilium* were included in the analysis.

Fig. 6. Maximum parsimony tree derived from analyses of the alignment of all codon positions from three genes, *rbcL*, *atpB*, and 18S rDNA. Monocots are represented by three genera with relatively primitive floral structure (*Tofieldia*, *Japonolirion*, *Lilium*). *Amborella* is revealed as the basal most angiosperm. Numbers above branches are jackknife support values found in analysis with ratchet algorithm (the method used by Soltis et al., 2000). Numbers below branches are bootstrap support values found by PAUP* (the method used by Goremykin et al., 2003).
Discussion

Amborella is not found to be sister to other angiosperms in analyses of the three gene data set performed using the same methods and almost the same taxon set as were used by Goremykin et al. (2003) (figs. 3, 5, 7). Rather, monocots and dicots are supported as two monophyletic groups sister to each other. These results are quite different from tree topologies inferred from analyses of the same three genes with much more extensive taxon sampling (Soltis et al., 2000). Amborella was supported as the basal most angiosperm in analyses by Soltis et al. (2000).

When Japonolirion, Tofieldia and Lilium are included into the tree gene data set, Amborella is revealed as sister to the rest of the angiosperms in our maximum parsimony and maximum likelihood analyses (figs. 6, 8). The basal position of Amborella, the paraphyly of dicots and the monophyly of monocots were revealed when grasses were excluded from the data set (fig. 6, 8) as well as when they were included together with Japonolirion, Tofieldia, and Lilium (tree not shown).

We suggest that the unusual tree topology published by Goremykin et al. (2003) reflects low taxon sampling rather than increased gene sampling. We consider that the data inferred from the analyses of 61 chloroplast genes (Goremykin et al., 2003) do not refute the idea of the basal position of Amborella. However, we realise that basal position of Amborella is nothing more than a hypothesis. Although current data on chloroplast genomes, in our opinion, are insufficient to refute this idea, it may be well refuted by further studies.

It is possible that the sister group relationships between monocots and dicots revealed by Goremykin et al. (2003) reflect low sampling of monocot taxa. In the study by Goremykin et al. (2003), only grasses are included in the analyses. The family Gramineae is usually treated as an advanced monocot group. In the analyses of the three gene data set, monocots are revealed as sister to dicots when they are represented only by grasses.

Fig. 7. Maximum likelihood tree derived from analyses of the alignment of all codon positions from three genes, rbcL, atpB, and 18S rDNA.

Monocots are represented only by grasses in this analysis (as well as in the analyses by Goremykin et al., 2003). TreePuzzle program is used in this analysis as well as in the analysis by Goremykin et al. (2003). Amborella is not revealed as the basal most angiosperm. Rather, as in analyses by Goremykin et al. (2003), Amborella is clustered with Calycanthus.
Fig. 8. Maximum likelihood tree derived from analyses of the alignment of all codon positions from three genes, \textit{rbcL}, \textit{atpB}, and 18S rDNA. Monocots are represented by the three genera with relatively primitive floral structure (\textit{Tofieldia}, \textit{Japonolirion}, \textit{Lilium}). TreePuzzle program is used in this analysis as well as in the analysis by Goremykin et al. (2003). \textit{Amborella} is revealed as the basal most angiosperm.

When monocots with relatively primitive floral features are added to the data matrix, monocots are revealed as derived from dicots. In maximum parsimony and maximum likelihood analyses, \textit{Amborella} is revealed as the basal most angiosperm. It is very important to analyse full chloroplast genomes of putatively basal monocots.

It is quite surprising how similar are the tree topologies inferred from neighbour-joining analyses of the 61-gene data set (fig. 1) and the 3-gene data set (fig. 3) for almost the same set of taxa. In our opinion, this implies that estimation of angiosperm phylogeny inferred from the 3-gene data set is relatively realistic. The data by Goremykin et al. .

Fig. 9. Two maximum likelihood trees derived from analyses of the alignments of all codon positions from three genes, \textit{rbcL}, \textit{atpB}, and 18S rDNA. MrBayes program is used in these analyses. Left: monocots are represented by grasses. Right: monocots are represented by genera with relatively primitive floral structure. In both cases \textit{Amborella} is revealed as the basal most angiosperm.
(2003) have seemingly demonstrated that analyses of full chloroplast genomes may not effect revolutionary changes in the general outline of angiosperm phylogeny.

The present study shows that high (even 100 %) bootstrap or jackknife support of a particular clade does not ultimately indicate a 'true' topology (see also, e. g., Bremer, 1994. For example, dicots are revealed as 100 %-supported monophyletic clade in some analyses. However, some other analyses strongly suggest that the putative monophyly of dicots represents an artefact caused by low taxon sampling. Even multiple increase of the character set cannot guarantee a realistic tree topology if taxon sampling is low.

The use of different algorithms of maximum likelihood analysis yielded quite different tree topologies when grasses were included as representatives of monocots. The Bayesian inference of phylogeny revealed a basal position for *Amborella* with 100 % posterior probability (fig. 9, left). Analysis of the same data matrix using TreePuzzle program revealed *Amborella* as sister to *Calycanthus*, although the posterior probability was 61 % only (fig. 7). In our opinion, this finding indicates that significant advantages of the maximum likelihood algorithm are implemented in the MrBayes program. Unfortunately, Goremykin et al. (2003) did not conducted Bayesian analysis of the 61 gene matrix.

Acknowledgements

The work is supported by grants from RFBR (project N 03-04-4883 la) and Scientific Schools program (projects N 1712.2003.4 and 1898.2003.4).

LITERATURE CITED


РЕЗЮМЕ

Большинство современных исследований, основанных на молекулярных данных, указывает на то, что эндемичный для Новой Кaledонии монотипный род Amborella (Amborellaceae) является сестринской группой по отношению к большей класте, объединяющей все прочие покрытосеменных растения (например, Qiu et al., 2000; Soltis et al., 2000; Zanis et al., 2002; APG II, 2003). Однако работы, в которых был сделан вывод о базальном положении Amborella, основывались на изучении относительно небольшого (по отношению к общему объему генома) числа участков ДНК. В. В. Горемыкиным с соавторами (2003), изучив последовательности 61 белок-кодирующего гена хлоропластного генома, сделали вывод, что Amborella формирует сестринскую группу по отношению к Calycanthus, а не занимает базальное положение на филогенетическом древе покрытосеменных растений. По нашему мнению, результат В. В. Горемыкина и соавт. (2003) свидетельствует о том, что в предыдущих работах, числом проанализированных генов, а с недостаточным числом таксонов (13), привлекаемых для анализа. Мы полагаем, что анализ, основанный на данных по 61 гену, не опровергает идею об основании покрытосеменных растений Amborella.

Одна из известных работ, в которой был сделан вывод о базальном положении Amborella (Soltis et al., 2000), основана на анализе данных по трем генам (rbcL, 18S рДНК, atpB) у 567 таксонов. Для того чтобы противостоять новым данным о родственных связях Amborella, мы проанализировали данные по трем генам (rbcL, 18S рДНК, atpB) для примерно того же набора из 13 таксонов, который был исследован В. В. Горемыкиным и соавт. (2003). В результате мы получили дерево, очень близкое к тем, которые были построено по данным о 61 гене. Следовательно, простого сокращения числа изучаемых видов без увеличения числа анализируемых генов достаточно для того, чтобы топология дерева изменилась самым существенным образом. Поэтому до тех пор, пока данные по 61 гену не будут проанализированы у большего числа видов, их нецелесообразно использовать для корректировки сложившихся на основе молекулярных данных представлений о филогенезе цветковых растений.