Molecular Phylogenetics and Evolution 59 (2011) 556–566

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

First insights into fern *matK* phylogeny

Li-Yaung Kuo [a,1], Fay-Wei Li [b,1,2], Wen-Liang Chiou [c], Chun-Neng Wang [a,b,*]

[a] Institute of Ecology and Evolutionary Biology, National Taiwan University, 1 Roosevelt Road, Sec. 4, Taipei 10617, Taiwan
[b] Department of Life Science, National Taiwan University, 1 Roosevelt Road, Sec. 4, Taipei 10617, Taiwan
[c] Division of Botanical Garden, Taiwan Forestry Research Institute, 53 Nan-Hai Road, Taipei 10066, Taiwan

ARTICLE INFO

Article history:
Received 18 June 2010
Revised 19 February 2011
Accepted 5 March 2011
Available online 12 March 2011

Keywords:
DNA barcoding
Homoplasy
matK
Phylogenetic performance
Substitution evenness

ABSTRACT

MatK, the only maturase gene in the land plant plastid genome, is a very popular phylogenetic marker that has been extensively applied in reconstructing angiosperm phylogeny. However, the use of *matK* in fern phylogeny is largely unknown, due to difficulties with amplification: ferns have lost the flanking *trnK* exons, typically the region used for designing stable priming sites. We developed primers that are either universal or lineage-specific that successfully amplify *matK* across all fern families. To evaluate whether *matK* is as powerful a phylogenetic marker in ferns as in angiosperms, we compared its sequence characteristics and phylogenetic performance to those of *rbcL* and *atpA*. Among these three genes, *matK* has the highest variability and substitution evenness, yet shows the least homoplasy. Most importantly, applying *matK* in fern phylogenetics better resolved relationships among families, especially within eupolypods I and II. Here we demonstrate the power of *matK* for fern phylogenetic reconstruction, as well as provide primers and extensive sequence data that will greatly facilitate future evolutionary studies of ferns.

1. Introduction

Molecular phylogeneticists are always searching for suitable loci with which to reconstruct the tree of life. Compared to nuclear genomes, organellar regions have been more broadly exploited for this purpose because they seldom undergo duplication, are usually uniparentally inherited, and are more easily amplified. The first mitochondrial and chloroplast (cp) loci used were coding regions because they typically evolve more slowly and have few indels, rendering their alignment unambiguous at high taxonomic ranks. However, because coding regions are functionally constrained, they tend to have slower evolutionary rates that usually limit their phylogenetic inferences. The *matK* gene, the only gene to encode for a splicing-associated maturase in the land plant cp genome, has helped overcome this obstacle: it is sufficiently fast evolving to be useful at a wide range of taxonomic depths, while still permitting unambiguous alignment. In angiosperms, for example, *matK* has considerable variation, and is even comparable to non-coding regions, yet with few alignment problems (Müller et al., 2006; Lohne et al., 2007; Hilu and Barthet, 2008; Hilu et al., 2008). Consequently, *matK* is frequently employed in phylogenetic analyses at various taxonomic levels of angiosperms (e.g., Hilu et al., 2003; Döring et al., 2007; Xie et al., 2010), and was even proposed as a DNA barcoding region for land plants in combination with *rbcL* (CBOL Plant Working Group, 2009).

In seed plants, *matK* is usually nested within the *trnK* intron, in a *trnK–matK–trnK* order, in the large single copy region of the chloroplast genome (Sugita et al., 1985; Neuhaus and Link, 1987; Hilu and Barthet, 2008). As a result, *matK* can be amplified by using primers targeted at the conserved *trnK* exons. Most leptosporangiate ferns, however, lack stable priming regions due to the loss of the *trnK* exons (Wolf et al., 2003; Duffy et al., 2009; Gao et al., 2009). The absence of suitable priming sites thus impaired the development of this locus in fern phylogenetics. The frequently used chloroplast loci including *rbcL*, *atpA*, and *atpB*, however, have been unable to unambiguously resolve family-level relationships (Schuettpelz et al., 2006; Schuettpelz and Pryer, 2007; Smith et al., 2006, 2008), especially in eupolypods II (Sano et al., 2000). Applying regions that can be aligned and yet evolve in a rapid manner, such as *matK*, to reconstruct ancient radiations is a more-efficient route than is combining multiple-conservative coding regions (Hilu et al., 2003).

In this study, we designed new primer sets aimed at the *matK* coding region and its 5′ upstream genes, including *trnK*, *chlB*, and *rps16*. We attempted to: (1) ascertain where the *trnK* loss event took place in the evolution of ferns; (2) characterize the sequence properties and quantify the phylogenetic performance of fern *matK*, and further compare it with *rbcL* and *atpA*; (3) evaluate

* Corresponding author at: Institute of Ecology and Evolutionary Biology, Department of Life Science, National Taiwan University, 1 Roosevelt Road, Sec. 4, Taipei 10617, Taiwan. Fax: +886 2 23673374.
E-mail addresses: salvinia@hotmail.com (L.-Y. Kuo), fayweili@gmail.com (F.-W. Li), botwang@ntu.edu.tw (C.-N. Wang).
[1] These authors contributed equally to this work.
[2] Present address: Department of Biology, Duke University, Durham, NC 27708, USA.

1055-7903/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2011.03.010

whether applying *matK* to phylogenetic analyses can bring new insights to the evolutionary history of ferns; and (4) discuss the utility of the fern *matK* gene in future phylogenetic and DNA barcoding studies.

2. Materials and methods

2.1. Taxonomic sampling, DNA extraction, and sequence amplification

We sampled 78 fern species from all 37 families (*sensu* Smith et al., 2006), with four seed plants and two lycophytes used as outgroups. DNA was extracted using a modified CTAB procedure (Wang et al., 2004) or a Plant Genomic DNA Mini Kit (GeneAid, Taipei, Taiwan) following the manufacturer's protocol.

We chose a subset of fern taxa (see Table 1) to screen for the presence of *trnK* exons, using the newly designed *trnK* primers, trnK 0118 and FernK r1 (Table 2; Fig. 1a). When *trnK* exons were present, the *matK* gene was within the *trnK–matK–trnK* amplicon and could be sequenced directly. For taxa without *trnK* exons, *matK* sequences were obtained via two separate amplifications. The first used a primer pair situated near the ends of the *matK* coding region (Fig. 1b). The primer pair, FERmatK fEDR and FERmatK rAGK, worked in most of ferns; however, in certain lineages specific primers had to be used (see Table 2 for details). The second amplification, to acquire the complete 5′ end of *matK*, required lineage-specific primers in the *matK* coding region (indicated by superscript d in Table 2; Fig. 1b) in combination with primers situated upstream in the *chlB* and *rps16* genes (indicated by superscript c in Table 2; Fig. 1b). Sequences from the two amplifications overlapped, and were combined to create a single contig. For species without available *rbcL* or *atpA* sequences on Genbank, these regions were also amplified and sequenced, using the published (Wolf et al., 1994; Schuettpelz et al., 2006) or newly designed primers (Table 2). PCR reactions were performed in a 15-μL volume, including >20 ng genomic DNA, 1× PCR buffer, 200 μM dNTP, 15 pmol of each primer, and 1 U polymerase (ProTaq, PROTECH, Taipei, Taiwan; or Phusion DNA polymerase, FINNZYMES, Espoo, Finland). For amplification of the complete 5′ end *matK* region (Fig. 1b), nested PCR was used by adding 1 μL of the first-round PCR product to second-round PCR with the same conditions described above.

2.2. Phylogenetic analyses

After ambiguous regions in the alignments were removed, maximum likelihood (ML) and Bayesian inference (BI) analyses were respectively carried out using Garli v0.96 (Zwickl, 2006) and MrBayes v3.1 (Ronquist and Huelsenbeck, 2003) for each of the seven datasets, including 3 one-gene datasets (*matK*, *rbcL*, and *atpA*); 3 two-gene datasets (*matK* + *rbcL*, *matK* + *atpA*, and *rbcL* + *atpA*); and a three-gene dataset (*matK* + *rbcL* + *atpA*).

For the ML analyses, a GTR + I + G model was used. The proportion of invariant sites and state frequencies were estimated by the program. The genthreshfortopoterm option was set to 20,000 while others followed the default settings. To calculate ML bootstrap support (BS) values, 1000 replicates were run under the same criteria. For the BI analyses, model selection and substitution parameters values were taken from Modeltest v3.7 (Posada and Crandall, 1998) as selected by the Akaike Information Criterion. Models and substitution parameters were individually assigned and unlinked across the gene partitions. Two simultaneous runs were carried out with four chains (of 10^6 generations each), in which each chain was sampled every 1000 generations. The first 25% of the sample was discarded as burn-in, and the rest used to calculate the majority-rule consensus tree.

2.3. Sequence characterization

For each gene, the nonsynonymous/synonymous substitution ratio (dN/dS) and variability (the proportion of variable sites) were calculated by PAML 4 (Yang, 2007) and Dnasp v4.50.3 (Rozas et al., 2003), respectively. MEGA v4.0 (Tamura et al., 2007) was used to infer the GC content and the transition/transversion ratio (ts/tv) for the three genes (*matK*, *rbcL*, and *atpA*). After excluding autoapomorphic insertions and incomplete sequence data at the ends of the alignments, MacClade v4.06 (Maddison and Maddison, 2003) was used to calculate the consistency index (CI), retention index (RI), and number of substitutions at each nucleotide site for all three genes based on the most likely tree from each one-gene dataset. A total of 1209 *matK*, 1176 *rbcL*, and 1491 *atpA* nucleotide sites were included. The value, $1-(\mathrm{RI}\times\mathrm{CI})$, was calculated for each nucleotide site to represent the degree of homoplasy at different codon positions (only nucleotide sites with more than one substitution event were included). To calculate the nucleotide substitution evenness (SE) of each codon position, the following formula based on the Shannon index (Shannon, 1948) was used, in which *pi* is the proportion of substitutions of a given nucleotide site to the total number of substitutions, and *S* is the total number of nucleotide sites:

$$\mathrm{SE}=\frac{-\sum_{i=1}^{s} pi\ln pi;}{\ln S}$$

3. Results

A total of 69 fern *matK* sequences were newly obtained in this study, spanning all recognized fern families (*sensu* Smith et al., 2006). Each sequence covered at least 4/5 of the gene length, and included a complete 5′ end (Table 1). Our *trnK* primers amplified *trnK-matK-trnK* amplicons from *Equisetum ramosissimum* (Equisetaceae), *Psilotum nudum* (Psilotaceae), *Ophioderma pendula* (Ophioglossaceae), *Botrychium formosanum* (Ophioglossaceae), *Danaea elliptica* (Marattiaceae), and *Osmunda japonica* (Osmundaceae), from which full-length *matK* sequences were obtained. For *matK* without flanking *trnK* exons, sequences were combined from two separate amplifications that gave two overlapping sequences (see Materials and Methods). We inadvertently sequenced the complete 3′ end of *Schizaea dichotoma*, due to the reverse primer, Sch matK rRDS, annealing to the 3′ downstream region of the expected priming site.

The log likelihoods of the maximum likelihood trees for the *matK* (**Supplementary material 1**) and three-gene (Figs. 2a and 2b) datasets were −57011.177 and −114004.4819. Maximum likelihood bootstrap support and Bayesian posterior probabilities for each branch are summarized in **Supplementary material 2**. No well-supported relationships were in conflict among the loci or phylogenetic methods.

Table 3 summarizes the sequence characteristics of *matK*, *rbcL*, and *atpA* genes in ferns, and compares these with angiosperms. The degree of homoplasy and the distribution of nucleotide substitutions for different codon positions are shown in Figs. 3 and 4, respectively.

4. Discussion

4.1. Loss of trnK gene in ferns

The loss of *trnK* exons (but with *matK* remaining intact) has been reported from several plant groups, including ferns (Wolf et al., 2003), *Selaginella* (Tsuji et al., 2007), *Epifagus* (Wolfe et al., 1992), and *Cuscuta* (Funk et al., 2007). Our results show that in