RESEARCH ARTICLE

Physicochemical evolution and positive selection of the gymnosperm matK proteins

DA CHENG HAO[1]*, JUN MU[1]*, SHI LIN CHEN[2] and PEI GEN XIAO[2]

[1]*Biotechnology Institute, College of Environment, Dalian Jiaotong University, Dalian 116028, People's Republic of China*
[2]*Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Beijing 100193, People's Republic of China*

Abstract

It is not clear whether matK evolves under Darwinian selection. In this study, the gymnosperm Taxaceae, Cephalotaxaceae and Pinaceae were used to illustrate the physicochemical evolution, molecular adaptation and evolutionary dynamics of gene divergence in matKs. *matK* sequences were amplified from 27 Taxaceae and 12 Cephalotaxaceae species. *matK* sequences of 19 Pinaceae species were retrieved from GenBank. The phylogenetic tree was generated using conceptual-translated amino acid sequences. Selective influences were investigated using standard d_N/d_S ratio methods and more sensitive techniques investigating the amino acid property changes resulting from nonsynonymous replacements in a phylogenetic context. Analyses revealed the presence of positive selection in matKs (N-terminal region, RT domain and domain X) of Taxaceae and Pinaceae, and found positive destabilizing selection in N-terminal region and RT domain of Cephalotaxaceae matK. Moreover, various amino acid properties were found to be influenced by destabilizing positive selection. Amino acid sites relating to these properties and to different secondary structures were found and have the potential to affect group II intron maturase function. Despite the evolutionary constraint on the rapidly evolving matK, this protein evolves under positive selection in gymnosperm. Several regions of matK have experienced molecular adaptation which fine-tunes maturase performance.

[Hao D. C., Jun M., Chen S. L. and Xiao P. G. 2010 Physicochemical evolution and positive selection of the gymnosperm matK proteins. *J. Genet.* **89**, 81–89]

Introduction

The *matK* gene is located in the large single-copy region of the chloroplast genome and between the 5′ and 3′ exons of *trnK* (tRNA-lysine), within group II intron. It is roughly 1500 bp in length, corresponding to 500 amino acids. The matK protein consists of three parts i.e., N-terminal region, reverse transcriptase (RT) domain in the middle, and domain X of C-terminus. The high rate of substitution in this gene has resulted in an increased number of parsimony informative sites and strong phylogenetic signals, making it useful to determine evolutionary histories at various taxonomic levels (e.g., Müller *et al*. 2006; Hao *et al*. 2008b). The abundant phylogenetic information derived from *matK* has made it an extremely valuable gene for DNA barcoding, systematic and evolutionary studies.

*For correspondence. E-mail: hao@djtu.edu.cn; mujun@djtu.edu.cn.

The matK protein is the only putative group II intron maturase that is encoded in the chloroplast genome. This putative function is based on the homology of a region in the carboxy terminus of matK to the conserved domain X of mitochondrial group II intron maturases (Neuhaus and Link 1987). Domain X of matK has a strong conserved sequence $SX_{3–6}TLAXKXK$, and most of the sequences have a large excess of basic over acidic amino acids (Mohr *et al*. 1993) and are mostly hydrophilic. Maturases are enzymes that catalyze non-autocatalytic intron removal from premature RNAs, such as RNA transcripts for the *trnK, trnA, trnI, rps12, rpl2*, and *atpF* genes (Vogel *et al*. 1999). The tRNA or protein products from these genes are required for normal chloroplast function, suggesting an indispensable function for matK in the chloroplast.

The *matK* gene is often chosen for phylogenetic reconstructions and it has been sequenced in thousands of plant

Keywords. physicochemical evolution; positive selection; chloroplast matK protein; Taxaceae; Cephalotaxaceae; Pinaceae.

species. Surprisingly, despite matK's physiological importance and abundance of sequence data, *matK* is generally used as strings of anonymous nucleotides, without regarding its functional evolution. It is noted that there is incongruence between the phylogenetic tress based on *matK* gene and that based on plant species morphology as well as between the former and nuclear ITS tree (e.g., Chaw *et al*. 2005; Rønsted *et al*. 2005). However, it is not clear whether natural selection is involved in such topology incongruence. Moreover, although *matK* pseudogene has been found in Valerianaceae (Hidalgo *et al*. 2004) and Corallorhizinae (Freudenstein and Senyo 2008), and purifying selection has been detected in the *matK* of non-photosynthetic Orobanchaceae (Young and dePamphilis 2000), little is known about *matK* evolution in other groups, e.g. in gymnosperms. Progression of *matK* towards a pseudogene state in some species is probably the reason for the lack of interest in the analysis of positive selection acting on it. Recently, Barthet and Hilu (2008) evaluated evolutionary constraints on *matK* using protein composition and tried to explain why this protein coding gene accommodates elevated rates of substitution and yet maintains functionality. Duffy *et al*. (2009) compared *matK* sequences of an intron-less fern clade to sequences from seed plants and ferns with the intron and found no significant differences in selection among lineages. It is thought that *matK* in ferns has maintained its ancient and generalized function in chloroplasts, even after the loss of its co-evolved group II intron. To gain deeper insight into the evolutionary pattern of matK of various plant groups, we for the first time detected positive selection of matK at both gene and protein levels using likelihood molecular phylogenetic analysis and quantitative physicochemical properties of amino acids, respectively. Using likelihood-based methods, a positive Darwinian selection amino acid site was found in both Taxaceae and Pinaceae, but not in the small families, Cephalotaxaceae. By evaluating positive selection in terms of the physicochemical amino acid properties which comprise protein phenotypes that selection at the molecular level may act upon, we identified positively selected residues, which may have experienced a shift in genic selection strategy, in all three regions of functional importance of all three families examined, except the domain X of Cephalotaxaceae.

Materials and methods

Taxon sampling, PCR amplification, sequencing

Sampling of Taxaceae and Cephalotaxaceae species, genomic DNA extraction, PCR amplification of *matK*, cloning and DNA sequencing were performed as previously described (Hao *et al*. 2008a,b, 2009a,b). Species' geographic origin of the sequenced material, voucher numbers and GenBank accession numbers of the sequences generated in this study, as well as those retrieved from GenBank, are given in table 1 of electronic supplementary material at http://www.ias.ac.in/jgenet/; 24 *matK* sequences were newly generated for this study. Most of the *matK* full-length sequences were PCR amplified using matf: ATGGATGAGTTCCAAAGATATGG and matr: TCATTTTTCTATTTGTTTATTATGTAT. trnK-3914F: GGGGTTGCTAACTCAACGG and trnK-2R: AACTAGTCGGATGGAGTAG were used to amplify *matK* from *Cephalotaxus harringtonia* var. *drupacea, C. fortunei* var. *alpina*, and *C. Koreana. matK* sequences from *C. hainanensis* and *C. griffithii* could not be amplified.

Phylogenetic analysis

DNA sequence and codon alignments were performed using RevTrans (Wernersson and Pedersen 2003, http://www.cbs.dtu.dk/services/RevTrans/) and CLUSTAL W2 (Larkin *et al*. 2007). The codon aligned Taxaceae + Cephalotaxaceae matK matrices comprised 531 (protein) and 1546 (nucleotide) positions, respectively. The best-fit model for the amino acid alignment was determined using ProtTest 1.2.6 (Abascal *et al*. 2005). DNA data were analysed using Modeltest 3.8 (Posada 2006) to find the best model of evolution for the data. Employing the Akaike information criterion (AIC), the model with the lowest AIC score was chosen. Maximum-likelihood (ML) analysis and bootstrapping were performed using RA×ML BlackBox (Stamatakis *et al*. 2008). The data sets were also analysed using MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003). Two independent runs with one cold and three heated Markov chains, each per analysis were performed simultaneously until the average standard deviation of split frequencies between the two runs dropped below 0.01. Analyses were run twice to check for consistency of results. We ran two simultaneous runs for 3×10^5 (protein) and 1.8×10^6 (nucleotide) generations, sampling trees every 100 (protein) and 500 (nucleotide) generations, respectively. Topology and branch-length informations were summarized in 50% majority rule consensus trees; samples obtained before stationarity of −ln likelihoods against generations had been reached were discarded as burn-in. The *matK* sequences of *Podocarpus*, Pinaceae, and *Fritillaria* were used as the reference for the rooted tree reconstruction.

Detection of positive selection

We tested for evidence of positive selection by comparing the nonsynonymous substitution rate (d_N) with the synonymous substitution rate (d_S). If a gene is evolving neutrally, $\omega = d_N/d_S$ is expected to be equal to one, whereas $\omega > 1$ is considered strong evidence that a gene experiences positive selection. We used several ML approaches to test for evidence of positive selection on matKs. The first approach, developed by Yang *et al*. (referred to as Yang models), involves comparisons of a neutral codon substitution model with ω constrained to be ≤ 1 to a selection model where a class of sites has $\omega > 1$ (Yang *et al*. 2000). As neutral models are nested within the corresponding selection models, a likelihood ratio test (LRT) can be used to compare them. The test statistic $-2\Delta l$nL (ΔlnL = the difference in log likelihoods of the two models) follows χ^2 distribution with de-

grees of freedom (df) equal to the difference in number of parameters between models. In the specific models implemented, ω varies between codons as a beta distribution (neutral: M7, M8a; selection: M8). We implemented models M7, M8a, and M8 with the codeml program in PAML4 (Yang 2007). Because Yang models are based on theoretical assumptions and ignore the empirical observation that distinct amino acids differ in their replacement rates, we also implemented MEC (mechanistic empirical combination) model (Doron-Faigenboim and Pupko 2007) that takes into account not only the transition–transversion bias and the nonsynonymous/synonymous ratio, but also the different amino acid replacement probabilities as specified in empirical amino acid matrices. Because the LRT is applicable when only two models are nested and thus is not suitable for comparing MEC and M8a models, the second-order AIC (AICc) was used for comparisons (Doron-Faigenboim and Pupko 2007). Those sites that are most likely to be in the positive selection class ($\omega > 1$) are identified as likely targets of selection.

Although the Yang models allow for variation in the nonsynonymous substitution rate, the synonymous rate is fixed across the sequence. Several methods for detecting positive selection that allow for variation in synonymous rate have been proposed, e.g., fixed effects methods and random effects methods. The fixed effect likelihood (FEL) method (Kosakovsky Pond and Frost 2005a) estimates ω on a site-by-site basis, uses ML estimation and treats shared parameters (branch lengths, tree topology and nucleotide substitution rates) as fixed. The random effects likelihood (REL) method is similar to the Yang model M3; however, both nonsynonymous and synonymous rates vary as gamma distributions with three rate classes (Kosakovsky Pond and Frost 2005a). The REL and FEL methods were implemented using the web interface DATAMONKEY (Kosakovsky Pond and Frost 2005b).

HYPHY models that allow d_N/d_S to vary among lineages were used to investigate whether selective pressure on TS and DBAT genes varies among lineages (Hao *et al*. 2009b). The genetic algorithm in HYPHY assigns four classes of d_N/d_S to lineages in a search for 'the best model' of lineage-specific evolution (Kosakovsky Pond and Frost 2005c), e.g., d_N/d_S = 10000, 0.933, 0.359, and 0 in Taxaceae, and d_N/d_S = 10000, 1.039, 0.565, and 0.233 in Pinaceae. This approach can identify lineages under positive selection without a *priori* hypothesis for lineage-specific evolution.

Recent methods have investigated selection in protein-coding genes further by addressing the type of positive selection detected (directional or nondirectional, stabilizing or destabilizing), the purifying selection, and how the identified selection affects the overall structure and function of the protein. For detecting selection in amino acid sequences we can look at the magnitudes of property change of nonsynonymous residues across a phylogeny. Amino acid substitutions have various effects on a protein depending on the difference in physicochemical properties and location in the protein structure. This approach facilitates differentiating between types of selective pressures and can detect positive and negative and stabilizing (selection that tends to maintain the overall biochemistry of the protein, despite a rate of change that exceeds the rate expected under conditions of chance) and destabilizing (selection that results in radical structural or functional shifts in local regions of the protein) selection and offer insights into the structural and functional consequences of the identified residues under selection (McClellan *et al*. 2005). We used the program TreeSAAP v3.2 (Woolley *et al*. 2003) to test for selection on amino acid properties within our matK data set. For each property examined, a range of possible 1-step changes as governed by the structure of the genetic code was determined and divided into eight magnitude categories of equal range, with lower categories indicating more conservative changes and higher categories denoting more radical changes. To construct an expected distribution of amino acid property change, each of the 9-nt changes in every codon of every DNA sequence within the data set was evaluated, with each nonsynonymous change assigned to one of the magnitude categories for each property independently. These property changes were then summed across the data set, constructing a set of relative frequencies of change for each of the eight magnitude categories to establish the null hypothesis under the assumption of neutral conditions (McClellan and McCracken 2001). If distributions of observed changes fail to fit the expected distributions based on goodness-of-fit scores and z-scores, the null hypothesis of neutrality is rejected. In terms of TreeSAAP analysis, positive destabilizing selection is defined as properties with significantly greater amino acid replacements than neutral expectations for magnitude categories 6, 7 and 8 (i.e., the three most radical property change categories). Within TreeSAAP, 31 amino acid properties are evaluated across a phylogeny using either the entire data set or a sliding window analysis. For our purposes, properties and magnitudes showing significantly more observed than expected numbers of changes at $P < 0.05$ level were first identified with an overall analysis of matK data. A sliding window analysis was then performed to investigate varying window sizes (10, 20, and 30 codons in width) to determine the range that maximizes the signal. The results of the sliding window analyses were used to identify regions in the protein that differ significantly from a nearly neutral model at a significance level of $P = 0.001$. Finally, we identified the particular amino acid residues within each region that contained positive destabilizing selection for each property. These residues might be of general importance to group II intron maturase function.

Results and discussion

Phylogenetic tree

Premature stop codons were not found in all *matK* sequences used, i.e., they are not pseudogenes. Conceptual translated amino acid sequences from consensus sequences of cloned

matK genes (15 Taxaceae and 8 Cephalotaxaceae taxa) as well as amino acid sequences acquired from GenBank were subjected to a phylogenetic analysis, and a Bayes/ML tree generated by MrBayes and RA×ML is shown in figure 1a. Cephalotaxaceae is the first-branching clade. *Cephalotaxus latifolia* is sister to other *Cephalotaxus* species. The matK of *C. wilsoniana* is closer to those of *C. koreana, C. harringtonia* cv. *fastigiata*, and *C. oliveri*, while the matK of *C. fortunei* is closer to that of *C. lanceolata*. The relationship among *C. mannii, C. fortunei* var. *alpina, C. harringtonia, C. sinensis*, and *C. harringtonia* var. *drupacea* is not resolved. *Torreya* and *Amentotaxus* form a well resolved clade. *Austrotaxus spicata* and *P. chienii* are basal to *Taxus* species. *Taxus globosa* is the first-branching species in *Taxus* clade, while *T. brevifolia* is the second one. The relationship among other *Taxus* matKs is largely unresolved, except that *T. wallichiana* is closer to *T. yunnanensis, T. floridana* is closer to *T. mairei*, and *T. sumatrana* is closer to *T. fuana*. This gene tree is significantly different from both the phylogenetic tree inferred from nuclear ITS (Hao *et al*. 2008b) and one generated by the combined analysis of five chloroplast DNA markers (Hao *et al*. 2008b). The topology of the matK tree may reflect: (i) cases where the same amino acid substitution occurred independently in more than one lineage; (ii) cases of the retention of plesiomorphic characters; and (iii) the possibility of incomplete lineage sorting.

In contrast, the congruence between Pinaceae matK protein tree and species tree (http://tolweb.org/Pinaceae) was revealed by Bayes and ML analysis (figure 1b). On the matK tree, *Pinus, Picea* and *Cathaya* formed a well-supported clade, which is closer to the clade consisting of *Pseudotsuga* and *Larix*. *Abies* is closer to *Keteleeria* than to *Tsuga* and *Pseudolarix*. *Cedrus* is the sister group of the remaining genera of the family Pinaceae. In order to evaluate how common positive selection in matK is among gymnosperm and where in the matK structure positive selection occur, we performed positive selection tests both at DNA and protein levels.

Positive selection tests at DNA (codon) level

Results from all five ML approaches for detecting selection indicated that a proportion of amino acid sites of matK in Taxaceae have evolved adaptively (table 1). Model MEC was best-fitting, as the log likelihood value was highest (−4205.48). The LRTs comparing Yang selection model M8 with neutral models (M7 and M8a) were significant (table 2). Compared to M8a, the MEC model had much higher log-likelihood value and much lower AICc score. The M8 model identified sites 118, 184, 495 and 497 as likely targets of positive selection (table 1; table 1 in electronic supplementary material), which were also identified by MEC model. Parameter estimates indicate that the positively selected class has a mean $\omega = 1.5$. For matK in Pinaceae, model M8 was best-fitting (log likelihood value −4712.81, table 3). LRT comparing M8 with M8a was significant (table 2), and compared to M8a, MEC model had higher log-likelihood value and lower AICc score. LRT comparing M8 with M7 was also significant ($P < 0.001$). Correspondingly, sites 138, 188, 300, and 410 identified by the M8 model as likely targets of positive selection were confirmed by MEC model (table 3; table 1 in electronic supplementary material).

The FEL method identified sites 97, 118, and 491 of Taxaceae matK as positively selected, while the REL method identified 15 sites (9, 54, 74, 86, 90, 97, 118, 122, 143, 184, 249, 327, 495, 497 and 508; table 1) as positively selected. It is noted that sites 118, 184, 495 and 497 were also identified by the M8 and MEC models. For matK in Pinaceae, both FEL and REL identified sites 49, 169, and 410 as positively selected sites, and 49 and 410 were also identified by M8 model. In contrast, no positively selected sites in Cephalotaxaceae were identified by any of the above methods.

In addition, we used a genetic algorithm approach (GA branch) that searches for an optimal model of lineage-specific evolution by assigning four unrestricted classes of d_N/d_S to lineages (Kosakovsky Pond and Frost 2005c). This approach allows an averaged model probability that d_N/d_S is greater than one along a specific lineage. Unlike branch site methods, GA branch does not need the user to select branches of interest to test, or test one branch at a time (which can lead to statisitcal instability or acceptance of poorly supported models), but rather mines the data for good-fitting models. In addition, inference based on multiple models (as opposed to a null-alternative pair) is more robust to model misspecification. For the *matK* locus in Taxaceae, the lineages T. *x media*, *T. sumatrana, Amentotaxus argotaenia*, and *Torreya californica* were placed into a d_N/d_S category of 10,000 (infinity; all substitutions along a given short branch are non-synonymous, figure 1a), with a model-averaged probability of 95.6%, 98.4%, 96.1%, and 99.7%, respectively. The 95% CIs for individual branch estimates of d_N/d_S were significantly different from one in these four cases. This variation might be mediated by the diversity in ecology and habitat that *Taxus* were exposed to as they radiated. However, it should be cautious when finding 'infinity' branches with only nonsynonymous as they could be due to the fact that there has been not enough divergence between the sequences and therefore no time to accumulate synonymous substitutions to get a reliable d_N/d_S estimate. For the *matK* locus in Pinaceae, the lineages *Abies holophylla* and *Keteleeria davidiana* were placed into a category d_N/d_S of 10,000 (figure 1b), with a model-averaged probability of 98.9% and 97.9%, respectively. The 95% CIs for individual branch estimates of d_N/d_S were significantly different from one in these two cases. On the contrary, for Cephalotaxaceae, although five branches were placed into a d_N/d_S category of 10,000 (data not shown), they failed to receive high model-averaged support for $d_N > d_S$.

The above findings at DNA (codon) level might correspond to an evolutionary demand for strategic biochemical and structural changes in some domains of the gymnosperm matK protein and an increase in selective constraints in

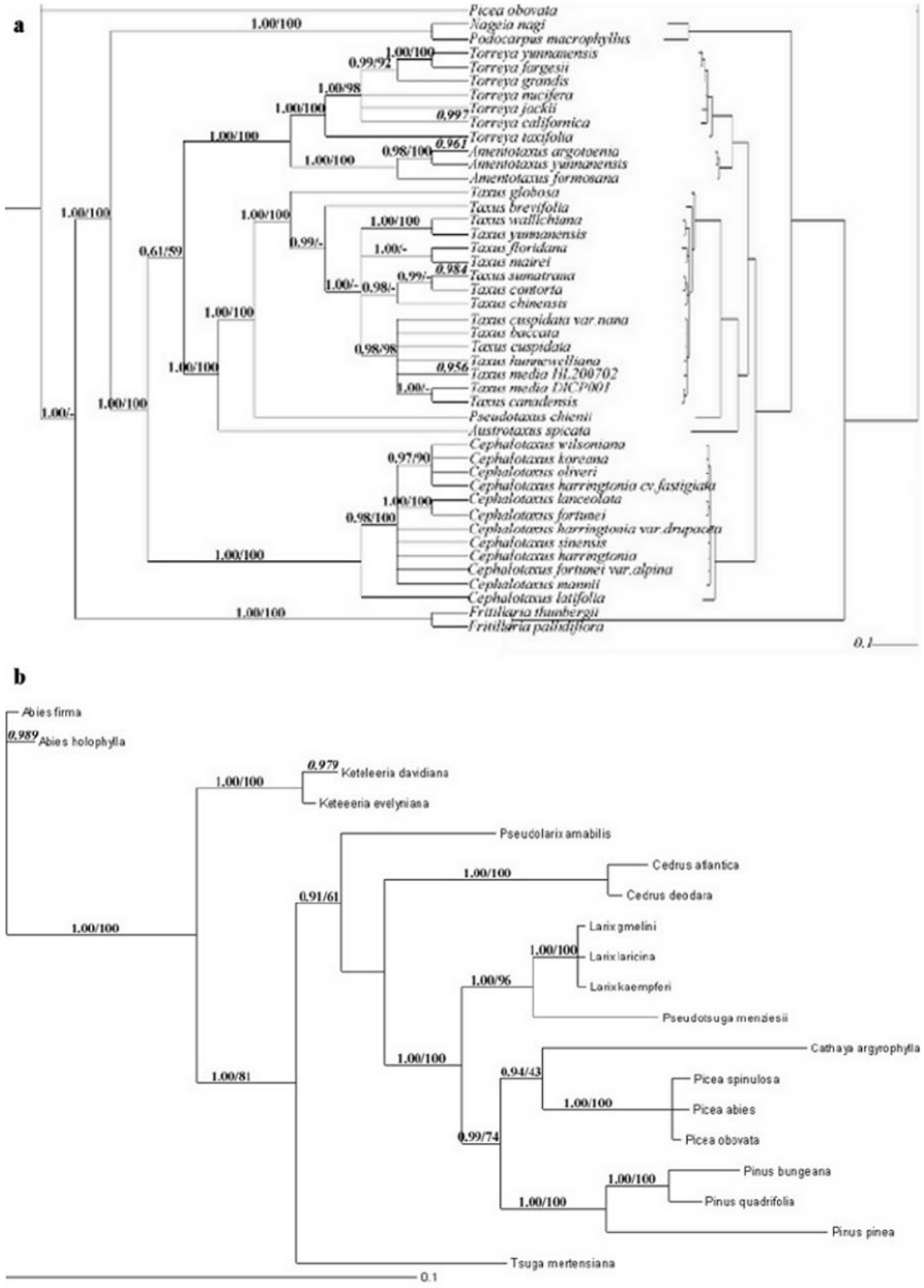

Figure 1. (a) Bayesian 50% majority rule consensus tree (3000 trees sampled; burn-in = 750 trees) inferred from the Taxaceae and Cephalotaxaceae matK amino acid alignment under the JTT model. Bayesian posterior probabilities (PPs) are given above branches, before slash. ML bootstrap values (BPs) are given after slash. Branch lengths (shown on the right; scale bar, expected number of substitutions per site) are proportional to the mean of the PPs of the branch lengths of the sampled trees. (b) Bayesian 50% majority rule consensus tree (2000 trees sampled; burn-in, 500 trees) inferred from the Pinaceae matK amino acid alignment under the JTT model. Bayesian PPs are given above branches, before slash. ML BPs are given after slash. Branch lengths (scale bar, expected number of substitutions per site) are proportional to the mean of the PPs of the branch lengths of the sampled trees. Italicized branch labells represent model averaged probabilities of $d_N/d_S > 1$ for the branch.

Table 1. Likelihood values and parameter estimates for the Taxaceae *matK* gene.

Model code	Log-likelihood	K	α	β	Parameter estimate	Positively selected sites[a]
M8[b] (beta and ω)	−4234.69	2.304	0.703	2.705	ω_s = 1.5 Prop(ω_s) = 0.277	118, 184, 495 and 497
M8a (null model)	−4238.67	2.199	0.117	2.705	ω_s set to 1 Prop(ω_s) = 0.5	Not allowed
M7 (beta)	-4239.23	2	0.430	0.554	ω_s = 1.5 Prop(ω_s) = 0	Not allowed
MEC[c]	−4205.48	–	0.218	1.869	rate (transition): 4.136 rate (transversion): 1.713 f : 0.618	118,184,392,495,497,508 and 509
FEL	–	–	–	–	15.16 subs/nucleotide	97, 118, 491
REL	–	–	–	–	12.62 subs/nucleotide	9, 54, 74, 86, 90, 97, 118, 122, 143, 184, 249, 327, 495, 497, 508

[a]Only sites with $d_N/d_S > 1$ where the 95% confidence interval is greater than 1 (i.e., the lower bound is larger than 1) are considered as significant. [b]M8, α and β are the shape parameters of the beta distribution. κ is the transition/transversion ratio. ω_s is the additional category representing positive selection. prop(ω_s) is the proportion of sites under selection. [c]MEC, f is the proportion of sites under no selection. Similar to PAML, the MEC model assumes a beta distribution with parameters α and β.

Table 2. Likelihood ratio statistics and AICc scores for tests of positive selection.

	M8 vs M8a(df=1)		MEC vs M8a		M8 vs M7(df=2)		
Plant family	Log-likelihood	*P*	Log-likelihood	AICc	Log-likelihood	−2ΔlnL	*P*
Taxaceae[a]	−4234.69/−4238.67	< 0.01	−4205.48/−4238.67	8421.0/8485.3	−4234.69/−4239.23	9.08	< 0.05
Cephalotaxaceae[b]	−2236.32/−2237.64	> 0.05	−2237/−2237.64	–	−2236.32/−2237.07	1.5	> 0.10
Pinaceae[c]	−4712.81/−4717.01	< 0.01	−4715.64/−4717.01	9441.3/9442.0	−4712.81/−4721.4	17.18	< 0.001

$-2\Delta\ln L = 2(\ln L_{\text{alternative hypothesis}} - \ln L_{\text{null hypothesis}})$, χ^2 distribution. [a]27 sequences, [b]12 sequences, [c]19 sequences.
$\text{AICc} = -2\log L + 2p\frac{N}{N-p-1}$. *L*, likelihood; *p*, no. of free parameters; *N*, the sequence length. The smaller the AICc value, the better the model explains the data.

Table 3. Likelihood values and parameter estimates for the Pinaceae *matK* gene.

Model code	Log-likelihood	k	α	β	Parameter estimate	Positively selected sites[a]
M8[b] (β & ω)	−4712.81	2.548	0.485	2.705	ω_s = 1.5 Prop(ω_s) = 0.302	49, 75, 120, 138, 188, 198, 212, 246, 281, 300, 311, 410 and 461
M8a (null model)	−4717.01	2.729	0.189	2.705	ω_s set to 1 Prop(ω_s) = 0.444	Not allowed
M7 (β)	−4721.4	2.558	0.312	0.401	ω_s = 1.5 Prop(ω_s) = 0	Not allowed
MEC[c]	−4715.64	–	0.115	1.491	rate (transition): 4.344 rate (transversion): 1.583 f: 0.554	138, 188, 300, 410 and 509
FEL	–	–	–	–	8.91 subs/nucleotide	49, 169 and 410
REL	–	–	–	–	6.86 subs/nucleotide	49, 169 and 410

[a]Only sites with $d_N/d_S > 1$ where the 95% confidence interval is greater than 1 (i.e., the lower bound is greater than 1) are considered as significant. [b]M8, α and β are the shape parameters of the beta distribution; κ, transition/transversion ratio; ω_s, additional category representing positive selection; prop(ω_s); proportion of sites under selection. [c]MEC, f is the proportion of sites under no selection. Similar to PAML, the MEC model assumes a beta distribution with parameters α and β.

others as habitat conditions influencing the adaptive landscape of optimal RNA splicing mechanisms significantly altered over a relatively short period of evolutionary time. We thus used a more sensitive technique to investigate the amino acid property changes resulting from nonsynonymous replacements in a historical (i.e. phylogenetic) context.

Positive selection tests at protein level

The d_N/d_S ratio analysis produced by codeml and MrBayes found no site in Cephalotaxaceae (but quite a few in Taxaceae) historically affected by positive selection using the criteria $d_N/d_S > 1.0$; no particular model was any better than any alternative using a likelihood ratio test. Selection models that implement d_N/d_S ratios as criteria for detecting selection are generally not sensitive enough to detect subtle molecular adaptations. It is, therefore, necessary to employ alternative criteria for the detection of positive selection among sites within generally conservative protein-coding genes. Although $d_N/d_S>1.0$ conditions most certainly indicate significant levels of historical positive selection, it is largely unreasonable to assume that Cephalotaxaceae matK did not adapt via selection. The evolutionary constraints on the rapidly evolving matK would preclude the obvious effects of positive selection by traditional criteria. However, if nonsynonymous substitutions are partitioned by the molecular-phenotypic effects of each positive selection for radical amino acid changes that may have a slower rate but occur more frequently than expected by chance may be more easily detected.

Significant physicochemical amino acid changes among residues in Cephalotaxaceae matK were identified by TreeSAAP, which compares the observed distribution of physicochemical changes inferred from a phylogenetic tree with an expected distribution based on the assumption of completely random amino acid replacement expected under the condition of selective neutrality. In Cephalotaxaceae, there are at least five radical amino acid property changes on the following sites, i.e., sites 52–59 and 61 of N-terminal region and sites 304–309 of RT domain (see table 2 in electronic supplementary material). The regions with the highest average number of changing properties per site are RT domain of Pinaceae and N-terminal region of Taxaceae (2.13 and 1.17 average changes per site, respectively), while there is no radical amino acid property change in domain X of Cephalotaxaceae (figure 2a). Further, there is at least one radical amino acid property change ($P < 0.001$) on most positively selected sites inferred from M8 and MEC models, e.g., in Taxaceae, site 184 is affected by positive-destabilizing selection for power to be at the C-terminal of an α-helix (α_c) and turn tendencies (Pt); in Pinaceae, site 300 is affected by positive-destabilizing selection for five structural properties (see table 2 in electronic supplementary material): Pt, power to be at the N-terminal of an α-helix (α_n), power to be at the middle of an α-helix (α_m), α-helical tendencies (P_α), and compressibility (K^0).

The amino acid properties found to be influenced by positive selection for destabilizing amino acid replacements are summarized by region in figure 2 and table 2 of electronic supplementary material. In Taxaceae, all three regions are affected by positive-destabilizing selection for isoelectric point (pHi), thermodynamic transfer hydrophobicity (Ht),

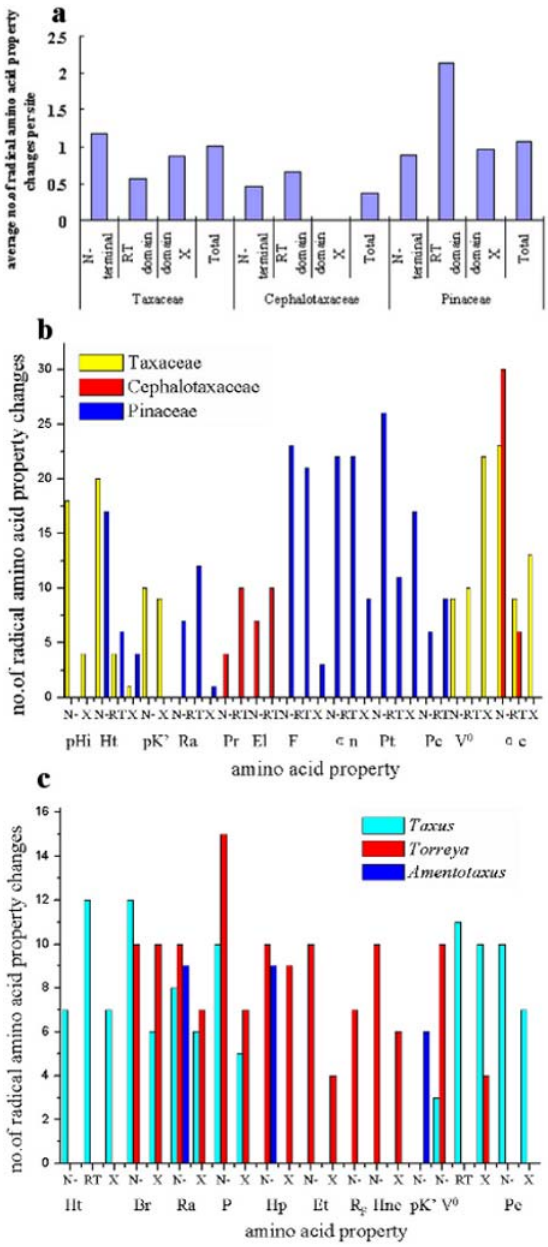

Figure 2. (a) Average number of radical amino acid property changes per site in three regions of gymnosperm matK. (b) Number of radical amino acid property changes in three regions of Taxaceae, Cephalotaxaceae, and Pinaceae matKs. In Taxaceae, five amino acid properties are detected to be under positive destabilizing selection: pHi, isoelectric point; V^0, partial specific volume; Ht, thermodynamic transfer hydrophobicity; pK′, equilibrium constant; α_c, power to be at the C-terminal of an α-helix. In Cephalotaxaceae, three amino acid properties are detected to be under positive destabilizing selection: E_l, long-range nonbonded energy; Pr, polar requirement; α_c. In Pinaceae, six amino acid properties are detected to be under positive destabilizing selection Ht; F, mean rms fluctuation displacement; α_n, power to be at the N-terminal of an α-helix; Pt, turn tendencies; Ra, solvent accessible reduction ratio; Pc, coil tendencies. N-, N-terminal region; RT, RT domain; X, domain X. (c) Number of radical amino acid property changes in three regions of *Taxus, Torreya* and *Amentotaxus* matKs. See text for details.

equilibrium constant (pK′), partial specific volume (V^0), and α_c, except no pHi and pK′ changes in RT domain (figure 2b). In Cephalotaxaceae, RT domain and domain X are affected by positive-destabilizing selection for α_c, long-range non-bonded energy (E_l), and polar requirement (Pr) (figure 2b). In contrast, in Pinaceae, all the three regions are affected by positive-destabilizing selection for Ht, solvent accessible reduction ratio (Ra), mean RMS fluctuation displacement (F), α_n, coil tendencies (Pc, except that domain X has no change in this structural property), and Pt (figure 2b). Further, we found there are six, eight, and three radical amino acid property changes in the Taxaceae genus *Taxus* (15 species), *Torreya* (seven species), and *Amentotaxus* (three species), respectively (figure 2c), and totally there are nine chemical property changes, compared to only two structural property changes, during the last 66 myr (Hao *et al*. 2008b). These chemical and conformational amino acid properties (Gromiha and Ponnuswamy 1993) may well be important to the overall optimization of matK function in gymnosperm and have been periodically adjusted during cladogenesis to maximize the biochemical effect of the spatial relationships between α-helices/β-sheets/loops and the primary functional amino acid residues that influence the RNA editing function of the intron maturase.

The exact RNA editing mechanism of matK is not clear, which precludes a clear discussion on the functional implications of the amino acid properties that are under selection (see figure 2a in electronic supplementary material). At the chemical level, moderate changes (categories 1 and 2) characterize more of the positive selection detected, while negative selection dominates the categories of radical changes; at the structural level, both radical changes and moderate changes characterize more of the negative selection detected; in Cephalotaxaceae, categories 1 and 7 changes were found to be statistically significant; chromatographic index is not changed in any plant family examined, regardless of category of amino acid property change. Other properties under positive destabilizing selection will interfere both at chemical and structural levels.

A crystal structure does not exist for any group II intron maturase. We, therefore, used secondary structure prediction to determine the distribution of radical and conservative amino acid property changes among three gymnosperm families. The conserved secondary structure of matK was predicted by two reliable and commonly used programs, JPRED (http://www.compbio.dundee.ac.uk/~www-jpred/) and GOR (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html). Interestingly, the distribution of radical and conservative amino acid property changes differs significantly among Taxaceae, Cephalotaxaceae, and Pinaceae (figure 2b in electronic supplementary material). In Taxaceae, more amino acid properties that are under positive destabilizing selection were found in β-sheets than in α-helices and loops (eight, six and four, respectively); in Cephalotaxaceae, seven amino acid properties that are under positive destabilizing selection were found in α-helices, compared to four and three in loops and β-sheets, respectively; in Pinaceae, more amino acid properties that are under positive destabilizing selection were found in loops than in helices and sheets (11, 4 and 3, respectively), indicating that positively selected sites located at the different secondary structures are not subject to random variation throughout the family as a result of structural and functional permissiveness on their locations. The findings of widespread positive selection in matK suggest that either selection still continues to improve performance of this ancient important enzyme, or that adaptive evolution in matK may reflect its fine-tuning to optimize its performance in various environmental conditions, with the latter being more probable. The adaptive evolution of chloroplast matK may have facilitated the successful radiation and diversification of gymnosperm species into different environments and habitats. The evidence of positive selection acting in functional regions of matK proteins provides the framework for future experimental characterization of the impact of specific mutations in the function, physiology, and interactions of matK protein involved in post-transcription splicing.

Positive selection in matK may facilitate horizontal inter-specific gene flow for chloroplast DNA, as spreading of adaptive mutations in matK may result in fixation of a single chloroplast haplotype in several occasionally hybridizing species, which may dramatically affect phylogeny reconstruction. We compared sums of bootstrap values between the trees reconstructed using all sites and the trees reconstructed using only neutrally evolving sites (positively selected sites excluded). The sum of bootstrap frequencies decreased by 7.2% in Taxaceae, while it increased by 10.9% in Pinaceae, suggesting that taking into account the presence of positive selection in *matK* may have major impact on the phylogenetic reconstructions. Previously, we detected cytonuclear discordance apparently caused by positive selection in matK and rbcL of Taxaceae and Cephalotaxaceae (Hao *et al*. 2008b). It might not be reliable to reconstruct phylogenetic and phylogenomic relations solely from chloroplast data in groups with putative inter-specific hybridization. Tests for the occurrence of positive selection and for the resemblance between chloroplast and nuclear phylogenies are essential for correct inference of species phylogenetic and phylogenomic relations.

Acknowledgements

This study is supported by Education Department of Liaoning Province (2009A120) and Start-up research fund (2008) of Dalian Jiaotong University.

References

Abascal F., Zardoya R. and Posada D. 2005 ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105.

Barthet M. M. and Hilu K. W. 2008 Evaluating evolutionary constraint on the rapidly evolving gene *matK* using protein composition. *J. Mol. Evol*. **66**, 85–97.

Chaw S. M., Walters T. W., Chang C. C., Hu S. H. and Chen S. H. 2005 A phylogeny of cycads (Cycadales) inferred from chloroplast *matK* gene, *trnK* intron, and nuclear rDNA ITS region. *Mol. Phylogenet. Evol*. **37**, 214–234.

Doron-Faigenboim A. and Pupko T. A. 2007 Combined empirical and mechanistic codon model. *Mol. Biol. Evol*. **24**, 388–397.

Duffy A. M., Kelchner S. A. and Wolf P. G. 2009 Conservation of selection on *matK* following an ancient loss of its flanking intron. *Gene* **438**, 17–25.

Freudenstein J. V. and Senyo D. M. 2008 Relationships and evolution of *matK* in a group of leafless orchids (Corallorhiza and Corallorhizinae; Orchidaceae: Epidendroideae). *Am. J. Bot*. **95**, 498–505.

Gromiha M. M. and Ponnuswamy P. K. 1993 Relationship between amino acid properties and protein compressibility. *J. Theo. Biol*. **165**, 87–100.

Hao D. C., Huang B. and Yang L. 2008a Phylogenetic relationships of the genus *Taxus* inferred from chloroplast intergenic spacer and nuclear coding DNA. *Biol. Pharm. Bull*. **31**, 260–265.

Hao D. C., Xiao P. G., Huang B., Ge G. B. and Yang L. 2008b Interspecific relationships and origins of Taxaceae and Cephalotaxaceae revealed by partitioned Bayesian analyses of chloroplast and nuclear DNA sequences. *Plant Syst. Evol*. **276**, 89–104.

Hao D. C., Huang B., Chen S. L. and Mu J. 2009a Evolution of the chloroplast *trnL–trnF* region in the gymnosperm lineages Taxaceae and Cephalotaxaceae. *Biochem. Genet*. **47**, 351–369.

Hao D. C., Yang L. and Huang B. 2009b Molecular evolution of paclitaxel biosynthetic genes *TS* and *DBAT* of *Taxus* species. *Genetica* **135**, 123–135.

Hidalgo O., Garnatje T., Susanna A. and Mathez J. 2004 Phylogeny of Valerianaceae based on *matK* and ITS markers, with reference to *matK* individual polymorphism. *Ann. Bot. (Lond).* **93**, 283–293.

Kosakovsky Pond S. L. and Frost S. D. 2005a Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol*. **22**, 1208–1222.

Kosakovsky Pond S. L. and Frost S. D. 2005b Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* **21**, 2531–2533.

Kosakovsky Pond S. L. and Frost S. D. 2005c A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol. Biol. Evol*. **22**, 478–485.

Larkin M. A., Blackshields G., Brown N. P., Chenna R., McGettigan P. A., McWilliam H. *et al*. 2007 ClustalW and ClustalX version 2. *Bioinformatics* **23**, 2947–2948.

McClellan D. A. and McCracken K. G. 2001 Estimating the influence of selection on the variable amino acid sites of the cytochrome b protein functional domains. *Mol. Biol. Evol*. **18**, 917–925.

McClellan D. A., Palfreyman E. J., Smith M. J., Moss J. L., Christensen R. G. and Sailsbery J. K. 2005 Physiocochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins. *Mol. Biol. Evol*. **22**, 437–455.

Mohr G., Perlman P. S. and Lambowitz A. M. 1993 Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acid Res*. **21**, 4991–4997.

Müller K. F., Borsch T. and Hilu K. W. 2006 Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting *matK, trnT-F* and *rbcL* in basal angiosperms. *Mol. Phylogenet. Evol*. **41**, 99–117.

Neuhaus H. and Link G. 1987 The chloroplast $tRNA^{Lys}$ (UUU) gene from mustard (*Sinapsis alba*) contains a class II intron potentially coding for a maturase-related polypeptide. *Curr. Genet*. **11**, 251–257.

Posada D. 2006 ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Res*. **34**, 700–703.

Ronquist F. and Huelsenbeck J. P. 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.

Rønsted N., Law S., Thornton H., Fay M. F. and Chase M. W. 2005 Molecular phylogenetic evidence for the monophyly of *Fritillaria* and *Lilium* (Liliaceae; Liliales) and the infrageneric classification of *Fritillaria. Mol. Phylogenet. Evol*. **35**, 509–527.

Stamatakis A., Hoover P. and Rougemont J. 2008 A rapid bootstrap algorithm for the RA×ML Web-Servers. *Syst. Biol*. **75**, 758–771.

Vogel J., Borner T. and Hess W. 1999 Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. *Nucleic Acids Res*. **27**, 3866–3874.

Wernersson R. and Pedersen A. G. 2003 RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res*. **31**, 3537–3539.

Woolley S., Johnson J., Smith M. J., Crandall K. A. and McClellan D. A. 2003 TreeSAAP: selection on amino acid properties using phylogenetic trees. *Bioinformatics* **19**, 671–672.

Yang Z. 2007 PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol*. **24**, 1586–1591.

Yang Z., Nielsen R. and Goldman N. 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449.

Young N. D. and dePamphilis C. W. 2000 Purifying selection detected in the plastid gene *matK* and flanking ribozyme regions within a group II intron of nonphotosynthetic plants. *Mol. Biol. Evol*. **17**, 1933–1941.

Received 5 August 2009; accepted 26 November 2009

Published on the Web: 8 April 2010