The Mitochondrial Genome of the Gymnosperm *Cycas taitungensis* Contains a Novel Family of Short Interspersed Elements, Bpu Sequences, and Abundant RNA Editing Sites

Shu-Miaw Chaw,* Arthur Chun-Chieh Shih,† Daryi Wang,* Yu-Wei Wu, †‡ Shu-Mei Liu,* and The-Yuan Chou§

*Research Center for Biodiversity, Academia Sinica, Taipei, Taiwan; †Institute of Information Science, Academia Sinica, Taipei, Taiwan; ‡Department of Informatics, Indiana University; and §Institute of Medical Biotechnology, Central Taiwan University of Science and Technology, Taichung City, Taiwan

The mtDNA of *Cycas taitungensis* is a circular molecule of 414,903 bp, making it 2- to 6-fold larger than the known mtDNAs of charophytes and bryophytes, but similar to the average of 7 elucidated angiosperm mtDNAs. It is characterized by abundant RNA editing sites (1,084), more than twice the number found in the angiosperm mtDNAs. The A + T content of *Cycas* mtDNA is 53.1%, the lowest among known land plants. About 5% of the *Cycas* mtDNA is composed of a novel family of mobile elements, which we designated as “Bpu sequences.” They share a consensus sequence of 36 bp with 2 terminal direct repeats (AAGG) and a recognition site for the Bpu 10I restriction endonuclease (CCTGAAGC). Comparison of the *Cycas* mtDNA with other plant mtDNAs revealed many new insights into the biology and evolution of land plant mtDNAs. For example, the noncoding sequences in mtDNAs have drastically expanded as land plants have evolved, with abrupt increases appearing in the bryophytes, and then in the seed plants. As a result, the genomic organizations of seed plant mtDNAs are much less compact than in other plants. Also, the *Cycas* mtDNA appears to have been exempted from the frequent gene loss observed in angiosperm mtDNAs. Similar to the angiosperms, the 3 *Cycas* genes nad1, nad2, and nad3 are disrupted by 5 group II intron squences, which have brought the genes into trans-splicing arrangements. The evolutionary origin and invasion/duplication mechanism of the Bpu sequences in *Cycas* mtDNA are hypothesized and discussed.

Introduction

Presently, the complete mitochondrial genomes (mtDNAs) of land plants are only known for 1 liverwort (*Marchantia polymorpha*, Oda et al. 1992), 1 moss (*Physcomitrella patens*, Terasawa et al. 2007), and 7 angiosperms, including 4 eudicots (*Arabidopsis thaliana*, Chaw et al. 1997; *Brassica napus*, Notsu et al. 2002) maize (*Zea mays*, Ogihara et al. 2005), and 3 monocot crops (rice [*Oryza sativa*], *Beta vulgaris*, Handa et al. 2003; wheat (*Triticum aestivum*, Okuda et al. 2002) maize (*Zea mays* subsp. *mays*, Clifton et al. 2004); wheat (*Triticum aestivum*, Oghihara et al. 2005) (table 1). The evolution of mtDNAs is more complex than that of plastid genomes (cpDNAs) because mtDNAs have horizontal gene transfer, uptake of plastid and nuclear DNA sequences, extreme size and rate variations, rapid gene rearrangements, and more trans-splicing exons and RNA editing sites (Berghorsten et al. 2004; Knoop 2004; Sugiyama et al. 2005; Wang et al. 2007). However, to date, no comparative mtDNA study has been performed in the 1,000 + species of gymnosperms, which are considered to be “inextricably related to the origins and early divergence of seed plants” (Chaw et al. 1997). In an effort to expand our understanding of mtDNA organization and evolution, we herein determined the *Cycas* mtDNA and examined the first gymnosperm mtDNA from Taitung of mtDNA organization and evolution, we herein determined (Chaw et al. 1997). In an effort to expand our understanding related to the origins and early divergence of seed plants’ ’

Materials and Methods

**Determination of the Complete mtDNA Sequence of *C. taitungensis***

Young leaves (less than 10-day-old) were collected from an 8-year-old *C. taitungensis* tree grown in the greenhouse of the Academia Sinica. Intact mitochondria were

Key words: mitochondrial genome, *Cycas*, RNA editing sites, repeats, mobile elements.

E-mail: smchaw@sinica.edu.tw.

doi:10.1093/molbev/msn099

Advance Access publication January 12, 2008

© 2008 The Authors. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
isolated using the method described by Kadowaki et al. (1996). Dnase I was used to digest any nuclear genome (nrDNA) contaminants. mtDNA was isolated according to a cetyltrimethylammonium bromide-based protocol (Stewart and Via 1993), sheared into random fragments of 2–3 kb with a Hydroshear device (Genomic Solutions Inc., Ann Arbor, MI), and then directly cloned into the EcoRV site of the pBluescriptSK vector to generate a shot-gun library. Shotgun clones were sequenced as previously described (Wang et al. 2007), except that each nucleotide had about 10/C2 coverage. Gaps were filled with specific primers designed based on the sequenced clones. The complete Cycas mtDNA sequence has been deposited in DNA Data Bank of Japan, under NCBI accession number BABI01000001.

Sequence Data Analysis

Annotation of Protein-Coding, rRNA, and tRNA Genes

A database search was carried out by using the National Center for Biotechnology Information’s (NCBI) Web-based Blast service (http://www.ncbi.nlm.nih.gov/BLAST/), and genes with e-values smaller than 0.001 were selected. The exact gene and exon boundaries were determined by alignment of homologous genes from available and annotated plant mtDNAs (table 1). Multiple sequence alignments were performed using the MAP2 Web service (http://deepc2 psi.iastate.edu/aat/map2/map2.html). Alignments of both nucleotide sequences and the translated amino acid sequences of the protein-coding genes were manually inspected. The tRNA genes were annotated using the tRNAscan-SE program (Lowe and Eddy 1997).

ORF Finding and Intron Identification

Identification of open reading frames (ORFs) was performed with the Web-based NCBI ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/). The standard genetic code was applied. The intron types were identified on the basis of their sequences and secondary structures (Michel and Ferat 1995). The predicted introns were also verified by alignment of their sequences with orthologous sequences previously elucidated from other species.

RNA Editing Sites

Putative RNA editing sites in protein-coding genes were predicted using the PREP-mt Web-based program (http://www.prep-mt.net/) (Mower 2005). To achieve a balanced trade-off between the number of false positive and false negative sites, the cutoff score (C-value) was set to 0.6 as suggested by the author. In addition, the top 4 genes with highest number of editing sites were verified by reverse transcription–polymerase chain reaction (RT–PCR) with gene-specific primers (supplementary table S7, Supplementary Material online).

Analysis of Repeats

Identification of repeats was carried out using the REPuter Web-based interface (http://bibiserv.techfak.uni-bielefeld.de/reputer/) (Kurtz et al. 2001). Both sequence directions (forward and reverse complement) were searched. The number of maximum computed repeats was set to 5,000. Overlapped repeating sequences were manually removed from each result. Information on tandem repeats was obtained using the tandem repeats finder (http://tandem.bu.edu/trf/trf.html) (Benson 1999). Searches for repeated regions and tandem repeats were carried out on all elucidated plant mtDNAs.

Phylogenetic Analysis

The 22 protein-coding genes common to the 11 sampled mtDNAs (atp1, atp4, atp6, atp8, atp9, ccmB, cob, cox1, cox2, cox3, nad1, nad2, nad3, nad4, nad4L, nad5, nad6, nad9, mttB, rps3, rps4, and rps12) were extracted for phylogenetic analysis. The sequences were separately

---

Table 1
MtDNAs of the 11 Plants Examined in This Study

<table>
<thead>
<tr>
<th>Classification</th>
<th>Scientific Name</th>
<th>Common Name</th>
<th>mtDNA Accession Number</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charophyta</td>
<td>Chara vulgaris</td>
<td>Green algae</td>
<td>AY267353</td>
<td>Turmel et al. (2003)</td>
</tr>
<tr>
<td>Land plants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bryophyta</td>
<td>Physcomitrella patens</td>
<td>Moss</td>
<td>NC_007945</td>
<td>Terasawa et al. (2007)</td>
</tr>
<tr>
<td>Marchantia polynympha</td>
<td>Liverwort</td>
<td>NC_001660</td>
<td>Oda et al. (1992)</td>
<td></td>
</tr>
<tr>
<td>Seed plants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gymnosperm</td>
<td>Cycas taitungensis</td>
<td>Taitung cycad</td>
<td>BABI01000001</td>
<td>Present study</td>
</tr>
<tr>
<td>Angiosperm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monocots</td>
<td>Triticum aestivum</td>
<td>Wheat</td>
<td>NC_007579</td>
<td>Ogihara et al. (2005)</td>
</tr>
<tr>
<td>Oryza sativa sp. Japonica</td>
<td>Rice</td>
<td>BA0000029</td>
<td>Notou et al. (2002)</td>
<td></td>
</tr>
<tr>
<td>Zea mays</td>
<td>Maize</td>
<td>NC_007982</td>
<td>Clifton et al. (2004)</td>
<td></td>
</tr>
<tr>
<td>Dicots</td>
<td>Beta vulgaris L.</td>
<td>Sugar beet</td>
<td>AP000396-7</td>
<td>Kubo et al. (2000)</td>
</tr>
<tr>
<td>Nicotiana tabacum</td>
<td>Tobacco</td>
<td>NC_006581</td>
<td>Sugiyama et al. (2005)</td>
<td></td>
</tr>
<tr>
<td>Brassica napus</td>
<td>Oilseed rape</td>
<td>AP006444</td>
<td>Handa (2003)</td>
<td></td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>Mouse-ear cress</td>
<td>NC_001284</td>
<td>Unseld et al. (1997)</td>
<td></td>
</tr>
</tbody>
</table>

Chaw et al.
aligned and then concatenated. Gaps and stop codons were removed manually. Divergence of nucleotide sequence between each pair of taxa was estimated in terms of the numbers of substitutions per synonymous (Ks) or non-synonymous (Ka) site, based on the Pamilo–Bianchi–Li method implemented in MEGA 3.0 program (Kumar et al. 2004). The Neighbor-Joining trees reconstructed with the Ka values and Ks values were rooted at Chara. The number of bootstrap replicates was set to 500. All phylogenetic analyses and tree reconstructions were performed using MEGA 3.0.

Results and Discussion
Evolution of mtDNA Organization in Land Plants

Characteristics of Cycas mtDNA and Insights into the Evolution of Land Plant mtDNAs

The complete mtDNA of C. taitungensis is a circular molecule of 414,903 bp (fig. 1). Table 2, which compares the main features of mtDNAs from a charophyte (Chara vulgaris), 2 bryophytes (Marchantia and Physcomitrella), and 7 angiosperms, shows that the Cycas mtDNA is about 6-, 2.2-, and 4.0-fold larger than those of Chara vulgaris molecule of 414,903 bp (fig. 1). Table 2, which compares Characteristics of Cycas mtDNA and Insights into the Evolution of mtDNA Organization in Land Plants

Results and Discussion
MEGA 3.0. analyses and tree reconstructions were performed using the number of bootstrap replicates was set to 500. All phylogenetic analyses and tree reconstructions were performed using MEGA 3.0.

Results and Discussion
Evolution of mtDNA Organization in Land Plants

Characteristics of Cycas mtDNA and Insights into the Evolution of Land Plant mtDNAs

The complete mtDNA of C. taitungensis is a circular molecule of 414,903 bp (fig. 1). Table 2, which compares the main features of mtDNAs from a charophyte (Chara vulgaris), 2 bryophytes (Marchantia and Physcomitrella), and 7 angiosperms, shows that the Cycas mtDNA is about 6-, 2.2-, and 4.0-fold larger than those of Chara, Marchantia, and Physcomitrella, respectively, but does not significantly differ from the average (414 ± 102 kb) of the previously elucidated angiosperm mtDNAs (P = 0.502). The A + T content of the Cycas mtDNA is 53.1%, the lowest among known algae and land plants. As further shown in table 2, the total numbers of protein- and tRNA-coding genes decrease from charophytes (39 and 26, respectively) to seed plants (29–40 and 17–27, respectively) (detailed in supplementary table S2, Supplementary Material online). In contrast, the numbers of rRNA gene species remain the same in all lineages, with the exception of obvious gene duplications in the Poaceae (grass family) and Beta lineages, in which some tRNA genes are also duplicated.

Table 2 also shows that noncoding sequences (spacers, introns, and pseudogenes) account for 89.9% of the Cycas mtDNA sequence, consistent with the proportions found in other angiosperm mtDNAs (89.4 ± 3.1%). As land plants evolved from charophycean green algae (Chara, 9.3%), the closest living relatives of land plants (Karol et al. 2001), the noncoding sequences have drastically expanded in the mtDNA, showing abrupt increases in the bryophytes and then in the seed plants. As a result, the genomic organizations of mtDNAs are much less compact in seed plants when compared with lower plants.

Repeated sequences that present in the genome as multiple copies comprise approximately 15.1% of the Cycas mtDNA (table 2). The repeats, very few of which are over 2-kb long, are evenly distributed across the genome and mainly occur in the noncoding regions, including the intergenic spacers and introns (fig. 1B). As shown in table 2, most mtDNAs of land plants (except for rice) have 2–5 times more repeated sequences than the mtDNA of Chara, whereas more than a quarter of the rice mtDNA consists of repeated sequences. Among the sampled plant mtDNAs, that of Cycas contains the highest percentage of tandem repeats (4.97%; fig. 2A, detailed in supplementary table S3 [Supplementary Material online]); these include a novel family of mobile elements, the Bpu sequences/elements (see below), which are mainly found in 3- or 4-copy arrays (fig. 2B). Ogihara et al. (2005) noted the presence of many repeats in wheat mtDNA and hypothesized that alternative physical structures may be adopted by wheat mtDNA. Therefore, it seems logical to hypothesize that alternative circular mtDNAs in various recombinant forms might coexist in Cycas cells in vivo.

No group I intron was detected in the mtDNA of Cycas or the other 7 previously elucidated angiosperm mtDNAs (table 2). This observation is consistent with Knoop’s (2004) speculation that group I introns were lost from the common ancestor of hornworts and tracheophytes. Although Cho et al. (1998) demonstrated that a group I intron located in the coxl gene is widespread among 48 angiosperm genera, the examined angiosperms represented an exceptionally patchy phylogenetic distribution and did not include any of our sampled 7 angiosperms. Furthermore, the authors estimated that the intron invaded the coxl gene by cross-species horizontal transfer over 1,000 times during angiosperm evolution and “is of entirely recent occurrence.”

In contrast, we found 20–25 group II introns in the mtDNAs of the examined land plants (table 2), nearly double the number found in those of their sister group, the charophytes (13). Similar to the angiosperms, the 3 genes nad1, nad2, and nad5 in Cycas are disrupted by 5 group II intron sequences, which have brought the genes into trans-splicing arrangements. Studies of Malek et al. (1997) and Malek and Knoop (1998) concluded that trans-spliced group II introns had evolved from formerly cis-spliced introns before the emergence of hornworts. Most recently, the evolving date was discovered to be even before the emergence of mosses (Groth-Malonek et al. 2005).

Gene Content and Evolution of Gene Loss in the mtDNAs of Land Plants

Our phylogenetic analysis using either nonsynonymous (Ka) or synonymous (Ks) substitutions of 22 mitochondrial protein-coding genes shared by the 11 studied plants generated identical tree topologies. Figure 3A shows the Neighbor-Joining trees reconstructed using the Ks and Ka values, respectively, with Chara being designated as the outgroup. The topologies of these 2 trees strongly indicate that, excluding the bryophytes Physcomitrella and Marchantia, the seed plants (including Cycas and the 7 angiosperms) and the angiosperms form nested monophyletic clades. Within the angiosperms, the monocots and eudicots constitute 2 distinct subclades. Noting that sisterhood relationship of Physcomitrella and Marchantia has to be treated with caution because the sampled seed plants share a long branch. In addition, a recent multigene analysis (Qiu et al. 2006) based on dense taxon sampling suggested that hornwort (including Marchantia) diverged before moss (including Physcomitrella). The Ka and Ks-derived branch lengths leading to Cycas are nearly equal (fig. 3A), whereas the Ks-based branch lengths for the other species are strikingly shorter than the corresponding Ka-based branches.
Statistical analysis using a $Z$-test further indicated that $K_a$ value for the Cycas branch is higher than the average $K_a$ of the 7 studied angiosperms ($P < 0.05$, $Z$-test). The elevated $K_a$ in Cycas suggests that a rapid evolution/divergence may have occurred in some of the protein-coding genes of the Cycas mtDNA. This result is consistent with the observation that some genes in the Cycas mtDNA contain abundant RNA editing sites (see below).

A total of 39 protein-coding genes were identified in the Cycas mtDNA, which is the highest gene number
<table>
<thead>
<tr>
<th>Features</th>
<th>Chara</th>
<th>Marchantia</th>
<th>Physcomitrella</th>
<th>Cycas</th>
<th>Triticum</th>
<th>Oryza</th>
<th>Zea</th>
<th>Beta</th>
<th>Nicotiana</th>
<th>Brassica</th>
<th>Arabidopsis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (bp)</td>
<td>67,737</td>
<td>186,609</td>
<td>105,340</td>
<td>414,903</td>
<td>452,528</td>
<td>490,520</td>
<td>569,630</td>
<td>368,799</td>
<td>430,597</td>
<td>221,853</td>
<td>366,924</td>
</tr>
<tr>
<td>A + T content (%)</td>
<td>59.1</td>
<td>57.6</td>
<td>59.4</td>
<td>53.1</td>
<td>55.7</td>
<td>56.2</td>
<td>56.1</td>
<td>56.1</td>
<td>55.0</td>
<td>54.8</td>
<td>55.2</td>
</tr>
<tr>
<td>Gene number* (protein-coding/rRNA)</td>
<td>[39/26/3]</td>
<td>[41/27/3]</td>
<td>[39/24/3]</td>
<td>[39/22/3]</td>
<td>[33/17/3]</td>
<td>[35/18/3]</td>
<td>[32/17/3b]</td>
<td>[29/18/32]</td>
<td>[36/21/3]</td>
<td>[33/17/3]</td>
<td>[32/17/3]</td>
</tr>
<tr>
<td>Coding sequences (%)</td>
<td>90.7</td>
<td>20.3</td>
<td>37.0</td>
<td>10.1</td>
<td>8.6</td>
<td>11.1</td>
<td>6.2</td>
<td>10.3</td>
<td>9.9</td>
<td>17.3</td>
<td>10.6</td>
</tr>
<tr>
<td>Repeat sequences (%)</td>
<td>3.2</td>
<td>10.1</td>
<td>11.6</td>
<td>15.1</td>
<td>10.1</td>
<td>28.8</td>
<td>11.4</td>
<td>12.5</td>
<td>10.8</td>
<td>5.5</td>
<td>7.8</td>
</tr>
<tr>
<td>Tandem repeat sequences (%)*</td>
<td>0.22</td>
<td>0.49</td>
<td>0.14</td>
<td>4.97</td>
<td>0.42</td>
<td>0.09</td>
<td>0.35</td>
<td>0.71</td>
<td>0.05</td>
<td>0.36</td>
<td>0.34</td>
</tr>
<tr>
<td>Introns</td>
<td>14</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Group I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cis-spliced</td>
<td>13</td>
<td>25</td>
<td>25</td>
<td>20</td>
<td>17</td>
<td>17</td>
<td>15</td>
<td>14</td>
<td>17</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Trans-spliced</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Cp-derived sequences (bp)</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>18,113</td>
<td>13,455</td>
<td>22,593</td>
<td>25,132</td>
<td>___</td>
<td>9,942</td>
<td>7,950</td>
<td>3,958</td>
</tr>
<tr>
<td>(%)</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>4.4</td>
<td>3.0</td>
<td>6.3</td>
<td>4.4</td>
<td>___</td>
<td>2.1</td>
<td>2.5</td>
<td>3.6</td>
</tr>
<tr>
<td>RNA editing sites</td>
<td>0</td>
<td>0</td>
<td>___</td>
<td>1,084e</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>370f</td>
<td>___</td>
<td>427</td>
</tr>
</tbody>
</table>

* Pseudogenes and unique ORFs (such as ORF222 reported in Handa 2003) were excluded. Upper: duplicate genes were counted only once. Lower within curly brackets: all duplicate genes were included.

* The plasmid-localized tRNA-Trp was included (Clifton et al. 2004).

* The REPuter program (http://bibiserv.techfak.uni-bielefeld.de/reputer) was used to obtain the estimates. All duplicate copies of repeats are included. Overlapped sequences were counted only once.

* The tandem repeats finder program (Benson 1999) was used to obtain the estimates.

* Data were not available from original papers.

* The PREP-mt program (Mower 2005) was used to obtain the estimates. The cutoff value of the reported score was set to 0.6.

* 370 C-to-U editing sites were identified in 28 gene/ORF transcripts of sugar beet (Kubo et al. 2000).
identified to date among the studied seed plant mtDNAs (fig. 3B). When the distributions of conserved genes in the mtDNA from the 11 sampled species (supplementary table S2, Supplementary Material online) are mapped to their respective branches in a maximum parsimony tree (fig. 3B; based on the data used in fig. 3A), it is possible then to estimate the time of loss of a particular gene. As shown in figure 3B, our analysis indicates that there have been at least 31 independent events of gene loss from all the land plant mtDNAs elucidated to date.

When a gene is missing from the mtDNA of a given species, it is generally believed that the original copy has been transferred to nucleus, where it functions through cytosolic protein synthesis followed by transit peptide–assisted import back to the mitochondria (Adams and Palmer 2003; Knoop 2004). Frequent gene losses, especially of the ribosomal protein genes (30 of 34), appear to have occurred after the divergence of the angiosperm lineages (approximately 150 MYA (Chaw et al. 2004). However, only 1 gene loss was observed in the Cycas lineage after it branched off from the common ancestor of angiosperms, approximately 300 MYA. Adams and Palmer (2003) suggested that angiosperm mtDNAs have experienced a recent evolutionary surge of loss and/or transfer of genes (primarily those encoding ribosomal proteins) to the nucleus. Our data give additional support to their contention but suggest that the mtDNA of Cycas appears to have been excluded from this surge. Our results suggest that the Cycas mtDNA tends to evolutionarily maintain its gene diversity and/or enjoys less gene transfer than other angiosperm mtDNAs (Selosse et al. 2001; Adams et al. 2002; Adams and Palmer 2003). Coincidently, among the some 40 published cpDNAs of seed plants, that of Cycas also undergoes the least gene loss (Wu et al. 2007, supplementary fig. 1 [Supplementary Material online]). Future work will be required to examine why the genomes of these 2 Cycas organelles appear to have frozen after the cycds divergence from angiosperms, especially in comparison to mtDNAs from other major gymnosperm clades, such as Ginkgo, gnetophytes, and pines.

A Novel Family of Short Interspersed Mitochondrial Elements

Numerous Short Interspersed Mitochondrial Elements, Termed Bpu Sequences, Are Present in the Cycas mtDNA

Sequence analysis revealed that numerous copies of a 36-nt repeat, herein designated as a “Bpu sequence/element,” are interspersed throughout the Cycas mtDNA (fig. 1B). Figure 4A shows the characteristic sequences of 500 Bpu elements having 0 to 4 mismatches. If up to a 7-nt mismatch to the dominant type is allowed, the total copy number of Bpu sequences increases to 512. The Bpu sequences feature 2 conserved terminal direct repeats (AAGG) and the recognition sequence for the restriction endonuclease, Bpu10I (CCTGAAGC; nt 15–21). These repeat elements/sequences do not appear to have coding potential.

Because Bpu elements are extremely short in their lengths and terminal repeats and contain 2 terminal direct repeats rather than the inverted repeats found in plant miniature inverted-repeat transposable elements (MITEs), they are here classified as short interspersed mitochondrial elements (SIMEs). This distinguishes them from the MITEs and short interspersed nuclear elements (SINEs, e.g., Alu DNA repeats in the genomes of primates) that have been extensively reported from the genomes of plants and animals (Feschotte et al. 2002). Based on the similarity between SIMEs and SINEs, however, we can hypothesize that Bpu elements are likely to be transposed through RNA without a requirement for reverse transcriptase.

The Bpu10I recognition site characteristic of the Bpu elements is highly conserved except at the very last base pair, that is, the 21st bp of Bpu sequence (fig. 4A, lower chart). In contrast, the 5′ terminal direct repeat and its downstream 6 nt (nt 5–9) tend to be more variable than the 3′ terminal direct repeat. Enigmatically, that elements can form a secondary structure (fig. 4B) with a predicted free energy of $-12.5$ (kcal/mole), as calculated by the MFOLD program (http://frontend.bioinfo.rpi.edu/zuker/ comments/FAQs.html). The significance of this secondary structure remains to be elucidated. Comparison of the Bpu
element insertion sites in the *Cycas* mtDNA with corresponding sites in the mtDNAs of other species reveals that the target sites for transpositions of Bpu elements are identical (within 1–2 mismatched base pair) to the 5’ terminal direct repeat (fig. 4C–F).

**Bpu Sequences Distinguish *Cycas* From Other Cycads and Seed Plants**

Bpu sequences are present exclusively in the noncoding regions of the *Cycas* mtDNA, that is, within the introns and intergenic spacers, with 1 exception: a Bpu sequence is present within the coding region of *rrn18*, which codes the 18S rRNA (fig. 4C). Among the many mitochondrial *rrn18* (mt-*rrn18*) genes available in GenBank, only those of *Cycas* (including *Cycas revoluta*, GenBank accession number: AB029356) and *Ginkgo biloba* (the only living species of the order Ginkgoales) have Bpu elements (fig. 4C). The Bpu insertion sites of the 2 *Cycas* taxa are orthologous (data not shown), whereas that of *Ginkgo* is different, indicating that the insertions of Bpu elements in *rrn18* did not occur in the common ancestor of cycads and ginkgo. In sequenced mtDNAs, the oldest cpDNA-derived sequences (also termed mtpt) cluster, *trnV*(uac)-*trnM*(cau)-*atpE*-*atpB*-rbcL, is reported to have existed since the common ancestor of seed plants (Wang et al. 2007; fig. 3). Bpu elements are present in the *Cycas* mtpt-*atpB* (fig. 4D) and mtpt-rbcL (fig. 4E) sequences but not in the corresponding sequences of the 7 angiosperm mtDNAs elucidated to date (Wang et al. 2007), suggesting that these Bpu insertions took place after the split of the *Cycas* lineage from the other seed plants.

We further examined the occurrence of Bpu elements in 1 of the other 2 cycad families, Zamiaceae (including 1 species from each of *Dioon, Macrozamia*, and *Zamia*) by sequencing the exons 1 and 2 of their *nad2* genes. Whereas the *nad2* of *Ginkgo* was similar to that of *Cycas* mtDNA, containing 5 Bpu elements, such elements are absent from the *nad2* genes of the 3 sampled Zamiaceae genera. However, we found 1, 2, and 6 Bpu elements in mitochondrial *nad5* of *Cycas panzhihuaensis* (GenBank accession number: AF43425) and mitochondrial *nad1* and cp *rps19-rpl16* spacers/intron of *C. revoluta* (GenBank accession number: AB029356).
FIG. 4.—Bpu sequences and examples of their insertion loci. Target sites and terminal repeats are underlined. (A) Upper: a dominant sequence based on 500 Bpu sequences (see text). Numbers along the abscissa indicate each nucleotide position within the prototype Bpu sequence. Recognition sites for the Bpu10I restriction enzyme are marked with asterisks. The ordinate scales each by the total bits of information multiplied by its relative occurrence at that position (Wasserman and Sandelin 2004). Lower: identities of Bpu sequences that differ from the dominant type by 0–4 bp. (B) Secondary structure of the dominant Bpu sequence. (C)–(F): lower cases, periods, and dashes denote mismatches, identical bps, and deletions, respectively, compared with the uppermost sequence. (C) Partial alignment of mt-rrn18 sequences from Cycas and Ginkgo. A Bpu sequence and a reversely complementary Bpu sequence are present in Ginkgo and Cycas, respectively. (D) Partial alignment of mtpt-atpB sequences extracted from...
AY354955, AY345867), respectively. These data seem to suggest that no Bpu sequence has successfully invaded the mtDNAs of Cycadales genera other than *Cycas*.

A Bpu-like sequence with a 1-bp insertion and 89% similarity to the dominant type in *Cycas* mtDNA was retrieved from the coding sequence of the mitochondrial *rps11* gene from the core eudicot, *Weigela hortensis*. Alignment of this Bpu-like sequence reveals that its predicted Bpu10I endonuclease recognition site differs from that of *Cycas* by 2 bp (gCTGAGt). Because this Bpu element-like recognition site differs from that of *Cycas*, we do not consider that this Bpu-like element shares a common origin with the Bpu sequences of *Cycas*. We further postulate that Bpu elements are likely absent from or very rare in angiosperms.

Surprisingly, the cpDNA of *Cycas* also contains 2 Bpu elements and each of them locates in the *petN-psbM* and *psbA-trnK* spacers, respectively (see Wu et al. 2007). Additionally, we also identified 1 Bpu sequence in the *atpB-rbcL* spacer of the cpDNA from *Pinus luchuensis* (GenBank accession number: DQ196799). However, no Bpu sequence has been detected in the cpDNAs of the 3 other Pinaceae genera and 3 gnetophyte orders we have sequenced to date (Chaw SM, Wu CS, Lai YT, Wang YN, Lin CP, Liu SM, unpublished data). Collectively, the available evidence inclines us to believe that the Bpu sequences have proliferated specifically in the mtDNA of *Cycas* (or the Cycadaeae). The sporadic occurrence of Bpu elements in the cpDNAs of *Cycas* and *Pinus* suggests that they are likely derived from nonhomologous recombination with DNA fragments that leaked out of mitochondria in the former case and via lateral transfer in the latter case.

The Fates of CpDNA-Derived Sequences (mtpts) in *Cycas* mtDNA

Table 2 shows a comparative analysis of mtpts among the 11 studied plant mtDNAs. The total percentage of mtpts in *Cycas* (4.4%) is relatively high compared with that in dicots (1.1–3.6%) but falls within the range seen among monocots (3.0–6.3%). We previously discovered that the frequency of mtpt transfer is positively correlated with variations in mtDNA size (coefficient value $r^2 = 0.47$) (Wang et al. 2007). Here, we report that the *Cycas* mtDNA contains 8 protein-coding genes, as well as 2 rRNA and 5 tRNA gene sequences originating from the cpDNA (fig. 1). However, frameshifts and indels within the protein-coding genes suggest that these *Cycas* mtpts have degenerated and are nonfunctional. In contrast, the 5 cpDNA-derived tRNAs in the *Cycas* mtDNA are able to fold into standard cloverleaf structures, as shown by tRNAseq-SE analysis (Lowe and Eddy 1997), and are thus likely to be functional. Previously, Sugiyama et al. (2005) used tRNAseq-SE to scan

---

*Cycas, Oryza, Arabidopsis,* and *Nicotiana.* Note that 2 Bpu sequence insertions are detected in *Cycas* but not in the other plants. (E) Partial alignment of mtpt-*rbcL* sequences extracted from *Oryza, Arabidopsis, Nicotiana,* and *Cycas.* The sequence from *Cycas* contains a Bpu-like sequence with a 5-bp insertion (shaded box) versus the dominant type. (F) Partial alignment of the mtpt-*atpE* and chloroplast *atpE* sequences of *Cycas.* Two reversely complementary Bpu sequences are found in the mtDNA. Note that the 5′ Bpu sequence is partly degenerated at its 3′ end.

---
the tobacco mtDNA and concluded that the 6 cp-derived tRNAs shared by angiosperms are functional. Because no mtpt was observed in *Chara*, *Marchantia*, or *Physcomitrella*, Wang et al. (2007) concluded that frequent DNA transfer from cpDNA to mtDNA has taken place no later than in the common ancestor of seed plants, approximately 300 MYA.

Abundant RNA Editing Sites in *Cycas* mtDNA

Among the land plant DNAs, *Cycas* has the most predicted RNA editing sites (1,084 sites; supplementary table S4 [Supplementary Material online]). It is commonly believed that RNA editing arose together with the first terrestrial plants (Steinhauser et al. 1999). By using the PREP-mt software (Mower 2005) with the cutoff score set to 0.6, 1,084 sites within the protein-coding genes of the *Cycas* mtDNA were predicted to be C-to-U RNA editing sites. This is more than double the number of predicted sites within the elucidated mtDNAs of other land plants (table 2).

If the cutoff score, which indicates the conservation degree of each editing site compared those found in the other published plant mtDNAs, is set to the most stringent criterion in PREP-mt (i.e., 1), the number of editing sites decreases to 738, which is still the largest number found among the land plant mtDNAs elucidated to date.

It is believed that RNA editing is essential for functional protein expression as it is required to modify amino acids or generate new start or stop codons (Hoch et al. 1991; Wintz and Hanson 1991; Kotera et al. 2005; Shikanai 2006). For this reason, the large number of RNA editing sites in the *Cycas* mtDNA may indicate higher complexity at the DNA level and formation of various transcripts through RNA editing, thus potentially reflecting rapid divergence in *Cycas*.

Although RNA editing sites are known to be sporadically distributed in the genomes of plant organelles/mitochondria from seed plants, the mechanisms underlying this distribution pattern are not yet known (Shikanai 2006;...
Mulligan et al. 2007). Supplementary table S4 (Supplementary Material online) shows that within the Cycas mtDNA, gene complex I (which includes 9 nad genes), is the most extensively edited gene category. Nad4 and nad5 show the first and second highest number of editing sites, followed by the cox1 gene of complex III. The least edited gene is rps10, which contains only 6 predicted sites even when the cutoff score is set to 0.6. Mulligan et al. (2007) interpreted the sporadic pattern as a result of lineage-specific loss of editing sites through retroconversion, which could remove adjacent editing sites by replacement with the edited sequences. The abundant RNA editing sites in the Cycas mtDNA, however, are mysterious and completely deviate from the sporadic patterns reported in other seed plant mtDNAs. Therefore, the mechanism of RNA editing in Cycas mtDNA may have a different evolutionary history from those of other seed plants.

Previous studies (Steinhauser et al. 1999; Lurin et al. 2004; Shikanai 2006) led to the proposal that variation or multiplication of trans-acting factors (e.g., members of the pentatriconopeptide repeat family) may allow rapid increases in the number of editing sites. This could possibly be correlated to the lineage-specific explosion of RNA editing sites observed in Cycas mtDNA. Furthermore, Handa’s (2003) conclusion that the evolutionary speed in RNA editing is higher at the level of gene regulation than at the primary gene sequence level appears to provide additional support for our speculation. To improve understanding of the mechanisms involved in abundant RNA editing sites in Cycas mtDNA, critical analysis of mtDNAs from ferns, conifers, and other cycads will be required.

In all plant lineages, RNA editing often alters the amino acid sequence (Adams and Palmer 2003). Supplementary table S5 (Supplementary Material online) summarizes the number of predicted preediting codons and potential edited codons in the Cycas mtDNA. The 2 most frequent edited codons are TCA (Ser) and TCT (Ser), which are predicted to be edited into TTA (Leu) and TTT (Phe), respectively. The least frequently edited codon is CAA (Gln), which would putatively be edited in only 2 cases, wherein the codon’s first position would be altered to yield the stop codon, TAA.

Supplementary table S6 (Supplementary Material online) further depicts the formation of 4 start (from ACG to ATG) and 7 stop codons (from CGA to TGA) through RNA editing. To verify the accuracy of the editing sites predicted using PREP-mt, partial transcripts of the cox1, atp1, atp6, and ccmB genes were experimentally assayed using RT-PCR. Comparison of our RT-PCR sequences with the editing sites predicted by PREP-mt indicated accuracy values of 97.66%, 99.13%, 97.91%, and 99.25% for cox1, atp1, atp6, and ccmB, respectively.

Our RT-PCR analysis additionally identified a U-to-C editing type in atp1; this was missed in the prediction because detection of the U-to-C editing type is limited when using PREP-mt (Mower 2005). Unfortunately, programs for predicting RNA editing sites in nonprotein-coding genes, such as ribosomal RNA and tRNA genes, and that of silent editing sites (which alter the mRNA but do not change the translated amino acid) are not yet available. We are currently examining these additional editing types and the locations of such sites in the Cycas mtDNA.

No Large Repeats in Gymnosperms but Some in Angiosperms

In the mtDNAs of angiosperms, long repeated sequences vary in size from 2,427 bp in rapeseed to 127,600 bp in rice, and they show no homology to each other, implying that repeated sequences were independently acquired during the evolution from Marchantia to angiosperms (Sugiyama et al. 2005). In Cycas mtDNA, we only detected a few long repeated sequences, all less than 1.5 kb in length and mostly composed of Bpu sequences. It has been demonstrated that genes are continuously transferred from the nuclear genome and cpDNA to mtDNA (Knoop 2004). However, the mechanisms underlying the emergence of large repeats are not yet fully understood.

Conclusions

The complete Cycas mtDNA shows a number of unprecedented features that are atypical of the mtDNAs previously elucidated for other land plants, including the lowest A + T content, the highest proportion of tandem repeat sequences (mainly Bpu sequences) and gene number found so far, abundant RNA editing sites, and an exceptionally elevated \( \kappa_p \) value for the protein-coding genes. The latter 2 features might be correlated. In comparison with the other known angiosperm organelle genomes, the cpDNA and mtDNA of Cycas have experienced the least gene loss. Peculiarly, the cpDNA of Cycas has the fewest RNA editing sites among land plants (Wu et al. 2007), whereas its mtDNA has the most. On the other hand, some characteristics are shared by mtDNAs of Cycas and angiosperms but not bryophytes; these include common gene transfers from cpDNA, similar ratios of noncoding sequences, lack of group I introns, relatively low numbers of tRNAs, and occurrence of trans-spliced group II introns. Nevertheless, the above-mentioned Cycas-specific features suggest that the Cycas mtDNA has a lineage-specific evolutionary history.

A novel family of SIMEs, designated as Bpu elements/sequences, represents another unique feature of the Cycadaeae lineage. Bpu elements are widely distributed throughout the noncoding regions of Cycas mtDNA (over 500 copies), with only 1 occurrence in a coding region (\textit{rrn} 18). In the Cycas mtDNA, the highly conserved nature of Bpu sequences and their association with Tai sequences as well as Worf760 (located in \textit{rps3.i2}) suggest that these sequences may have collectively originated from an ancestral retrotransposon. Further investigation into the origin, proliferation, and evolution of Bpu sequences will be desirable to help clarify the peculiar features of Cycas mtDNA.

Supplementary Material

Supplementary tables S1–S7 and figure 1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by joint research grants from the Research Center for Biodiversity, Academia Sinica, to
S.-M.C. and D.W., and partially by National Science Council grant to S.-M.C. (94-2311-B001-059) and by a grant from the Institute of Information Science, Academia Sinica, to A.C.-C.S. We thank Chiao-Lei Cheng for sequencing the RT-PCR products. We are grateful for the 3 anonymous reviewers, who provided critical comments and valuable suggestions.

Literature Cited


William Martin, Associate Editor

Accepted January 6, 2008