Phylogenetic analysis of the internal transcribed spacer (ITS) region in Menyanthaceae using predicted secondary structure

Nicholas P. Tippery *, Donald H. Les

Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N Eagleville Road U-3043, Storrs, CT 06269, USA

**A R T I C L E   I N F O**

Article history:
Received 11 March 2008
Revised 26 July 2008
Accepted 30 July 2008
Available online 6 August 2008

Keywords:
Internal transcribed spacer
RNA secondary structure
Phylogeny
 Parsimony
Likelihood
Menyanthaceae

**A B S T R A C T**

Sequences of the nuclear internal transcribed spacer (ITS) regions ITS1 and ITS2 have been used widely in molecular phylogenetic studies because of their relatively high variability and facility of amplification. For phylogenetic applications, most researchers use sequence alignments that are based on nucleotide similarity. However, confidence in the alignment often deteriorates at taxonomic levels above genus, due to increasing variability among sequences. Like ribosomal RNA (rRNA) and other RNA molecules, the ITS transcripts consist in part of conserved secondary structures ('stems' and 'loops') that can be predicted by mathematical algorithm. Researchers have long considered the evolutionary conservation of rRNA secondary structure, but until recently few phylogenetic analyses of the ITS regions specifically incorporated structural data. We outline a novel method by which to derive additional phylogenetic data from ITS secondary structure in order to evaluate support for relationships at higher taxonomic levels. To illustrate the method, we describe an example from the plant family Menyanthaceae. Using predicted ITS secondary structure data, we obtained a well-resolved and moderately supported phylogeny, in which most topological relationships were congruent with the tree constructed using ITS nucleotide sequence data. Furthermore, the explicit encoding of ITS structural data in a phylogenetic framework allowed for the reconstruction of putative ancestral states and structural evolution in the functional but highly variable ITS region.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

It has long been known that nucleotide sequences are constrained by the functions of the end products they encode, evidence for which includes the unequal accumulation of synonymous vs. non-synonymous mutations in protein-coding regions of DNA (Zuckerkandl and Pauling, 1965). Consequently, it is not surprising that phylogenetic models perform better when they account for unequal rates of substitution among sites (Buckley et al., 2001) and for factors that affect sequence conservation or variability (e.g., Powell and Moriyama, 1997). In ribosomal RNA (rRNA) molecules, which accomplish their function through complex secondary (and higher-order) structures determined by complementary base-pairing of linear RNA transcripts, observations of sequence conservation and compensatory nucleotide changes have facilitated the elucidation of conserved secondary structure (Gutell et al., 2002). Phylogenetic methods that account for functional constraints on RNA structure include down-weighting of sites in paired 'stem' regions (Wheeler and Honeycutt, 1988; Steele et al., 1991; Dixon and Hillis, 1993), alignment of multiple sequences according to secondary structure (Kjer, 1995; Gottschling et al., 2001; Goertzen et al., 2003), and linkage of complementarily paired sites in likelihood analyses (Schöniger and von Haeseler, 1994; Kjer, 2004).

The nuclear internal transcribed spacer (ITS) regions, which are interspersed among the rRNA genes, have been sequenced widely because of their relatively high variability and facility of amplification. The ITS regions are indispensable in the production of mature rRNA molecules because they enable their own excision from the RNA transcript (Joseph et al., 1999; Venema and Tollervey, 1999; Côté et al., 2002). Although several strictly conserved nucleotide sequence motifs have been identified in ITS1 and ITS2 (Liu and Schardl, 1994; Mai and Coleman, 1997), many ITS molecular interactions depend more upon a functionally conserved secondary structure than on the specific nucleotide sequence itself (van Nues et al., 1994, 1995; Joseph et al., 1999; Michot et al., 1999). Predicted ITS secondary structures that have been modeled by minimum free energy optimization (Zuker, 1989; Hofacker et al., 2002) are remarkably similar between distantly related taxa (e.g., algae and angiosperms), with respect to both their overall structure and the positions of certain conserved motifs (Hershkovitz and Lewis, 1996; Hershkovitz and Zimmer, 1996; Mai and Coleman, 1997; Coleman et al., 1998; Schultz et al., 2005; Wolf et al., 2005). In a detailed phylogenetic survey within Asteraceae (Magnoliophyta), Goertzen et al. (2003) were able to resolve seven...
conserved subregions of ITS (three in ITS1 and four in ITS2) that were similar to the subregions reported for other green plants and even more distantly related taxa. Conservation of both structure and sequence in ITS thus represents a broadly observed phenomenon, to which phylogenetic methods would be well suited.

Several studies have incorporated the phylogenetic conservation of sequence and structure into analyses of the ITS regions. Conserved oligonucleotide motifs have been used to anchor multiple sequence alignments (Coleman, 2003; Chen et al., 2004), and predicted secondary structures have helped to discriminate between variable and conserved nucleotide positions (Goertzen et al., 2003; Fougère-Danezan et al., 2007; Krüger and Gargas, 2008). In a method relying more explicitly on predicted secondary structure, Wang et al. (2007) conducted a phylogenetic analysis of ITS1 sequences using a distance matrix that reflected complementary base pair changes between pairs of taxa, which built upon similar work that had been done using rRNA coding regions (Billoud et al., 2000; Caetano-Anollés 2002). With ITS nucleotide sequences becoming increasingly abundant over a broad range of taxa, resources such as the ITS2 database (Schultz et al., 2006) have provided a more comprehensive understanding of ITS structure conservation, and methods that incorporate both nucleotide and structure conservation are being used.

Fig. 1. Predicted secondary structure for the Nymphoides cordata ITS1 and ITS2 regions (the intervening 5.8S rRNA has been omitted). Stem subregions correspond to conserved structures identified by Goertzen et al. (2003) for Asteraceae. Illustration produced using the program XRNA (B. Weiser and H. Noller, University of California, Santa Cruz).

Fig. 2. Summary of the character encoding method. Example predicted secondary structures for the 1A subregion of ITS1 are given at left for three taxa. Conserved stem portions (i.e., found in all example taxa) are bracketed in the predicted structures and in the encoded and aligned data at right. For the nucleotide data, the 5’ half of the sequence is given from left to right, while the 3’ half appears from right to left; thus, nucleotides that form a complementary base pair appear one directly above the other. For the pairwise interaction data, 1 indicates complementarily pairing nucleotides, 0 indicates non-pairing nucleotides, and a dash (−) represents an alignment gap. For additional visual clarity, mismatch and unpaired nucleotides are depicted in lower case, while complementarily pairing nucleotides are in upper case; gaps opposite unpaired nucleotides are depicted by a period (·).
structural data have allowed researchers to conduct phylogeny estimation on multiple sequences (Höchsmann et al., 2003, 2004; Siebert and Backofen, 2005; Seibel et al., 2006).

Although many studies have combined RNA secondary structure prediction and phylogeny estimation, most phylogenetic analyses that incorporate ITS structure remain dependent upon the expectation of site-specific nucleotide homology (i.e., that strings of nucleotides with a shared evolutionary background consistently comprise similar RNA structures). Studies that compare rRNA secondary structures among divergent taxonomic groups rely on the relatively high degree of sequence conservation in rRNA genes, which allows for confident assessment of site homology and identification of compensatory base changes (Hickson et al., 1996; Guettel et al., 2002). In the highly variable ITS regions, however, reliable homology assessment with respect to both nucleotide position and RNA secondary structure requires dense sampling among closely related taxa, where differences in secondary structure and component nucleotides are finer (Goertzen et al., 2003). At an intermediate taxonomic level (approximately that of genus), rRNA gene sequences are too invariant to be informative, and the variation within ITS nucleotide sequences makes them increasingly difficult to align (Coleman, 2003; Goertzen et al., 2003). ITS secondary structures are nonetheless comparable and phylogenetically informative among higher taxonomic groups, where similar predicted ITS secondary structures often are composed of highly divergent strings of nucleotides (e.g., Hershkovitz and Zimmer, 1996). Careful examination of structural changes at lower taxonomic levels (e.g., among species) may reveal the mechanisms by which secondary structures

### Table 1

<table>
<thead>
<tr>
<th>Data type</th>
<th># Characters</th>
<th># Parsimony informative (%)</th>
<th>( \theta_i )</th>
<th># Trees (MP)</th>
<th>Tree length (MP)</th>
<th>CI (MP)</th>
<th>RI (MP)</th>
<th>CIexc (MP)</th>
<th>lnL (BI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITS1</td>
<td>353</td>
<td>151 (43)</td>
<td>-0.65</td>
<td>6</td>
<td>742</td>
<td>0.74</td>
<td>0.85</td>
<td>0.69</td>
<td>-4672</td>
</tr>
<tr>
<td>5.8S</td>
<td>168</td>
<td>11 (07)</td>
<td>-0.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITS2</td>
<td>290</td>
<td>129 (44)</td>
<td>-0.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total nucleotide</td>
<td>811</td>
<td>291 (36)</td>
<td>-0.75</td>
<td>6</td>
<td>742</td>
<td>0.74</td>
<td>0.85</td>
<td>0.69</td>
<td>-4672</td>
</tr>
<tr>
<td>1A</td>
<td>38</td>
<td>18 (47)</td>
<td>-0.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B</td>
<td>47</td>
<td>22 (47)</td>
<td>-0.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1C</td>
<td>8</td>
<td>2 (25)</td>
<td>-0.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2A</td>
<td>52</td>
<td>22 (42)</td>
<td>-0.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2B</td>
<td>22</td>
<td>5 (23)</td>
<td>-0.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2C</td>
<td>62</td>
<td>26 (42)</td>
<td>-0.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2D</td>
<td>13</td>
<td>4 (31)</td>
<td>-2.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total structural</td>
<td>243</td>
<td>99 (41)</td>
<td>-0.72</td>
<td>1</td>
<td>277</td>
<td>0.58</td>
<td>0.75</td>
<td>0.52</td>
<td>-1146</td>
</tr>
</tbody>
</table>

CI, consistency index; RI, retention index; CIexc, CI excluding uninformative characters; lnL, natural log likelihood (harmonic mean).

**Fig. 3.** Strict consensus phylogenetic tree (left) derived from maximum parsimony analysis of ITS nucleotide sequences (including the 5.8S rDNA gene). Percent parsimony bootstrap support (above) and Bayesian posterior probability (below) values are given for each node. A phylogram for one of six most-parsimonious trees, with branch lengths, is depicted at right.
are conserved in more divergent lineages (e.g., among genera and higher levels) and uncover additional functional dependence upon conserved ITS secondary structures.

In this paper, we introduce a method to derive phylogenetically useful characters from RNA secondary structures predicted for the ITS1 and ITS2 regions. Focusing on conserved structures that have been reported previously for Asteraceae, we examined ITS sequences in the related family Menyanthaceae. Our accessory data matrix, consisting of pairwise nucleotide interaction data, constituted a novel set of characters that were relatively independent of the nucleotide sequences on which they were based. The accessory data accounted for nucleotide substitutions and insertions or deletions (indels) that altered the predicted ITS secondary structure, without being limited by the expectation of nucleotide position homology. Furthermore, analyzing the structural data in a phylogenetic context enabled us explicitly to reconstruct character states and evolutionary transitions among hypothesized ancestral taxa.

2. Materials and methods

2.1. Taxon sampling

Complete ITS nucleotide sequences (including ITS1, ITS2, and the 5.8S rRNA gene) were obtained from prior phylogenetic work in Menyanthaceae (Tippery et al., 2008; GenBank Accession Nos. EF173022–EF173059 and EU257161–EU257172). Twenty-four taxa were sampled out of 60–70 spp. in the family, including the three monotypic genera (Liparophyllum, Menyanthes and Nephrophyllum), eight species of Nymphoides (40–50 spp.), and 13 species of Villarsia (18 spp.).

2.2. Secondary structure prediction

Menyanthaceae are closely related to Asteraceae (Lundberg and Bremer, 2003), for which conserved secondary structure features have been determined previously (Goertzen et al., 2003). Conserved subregions identified by Goertzen et al. (2003) for Asteraceae, designated 1A, 1B, and 1C for ITS1, and 2A, 2B, 2C, and 2D for ITS2, provided a framework for secondary structure modeling (Fig. 1). For a particular subregion (e.g., 1A), sequences were trimmed to within 10 nucleotides of the predicted Asteraceae structure at each end. Trimmed sequences were input into Quikfold on the DINA melt Server (Zuker, 2003; Markham and Zuker, 2005) to determine putative secondary structure, using the following parameters: linear sequence, RNA version 2.3 energy rules, 20°C. Although our nucleotide sequences were obtained from amplification and sequencing of nuclear DNA, they were treated as RNA transcripts for the purpose of modeling. The top five percent of optimal and suboptimal structures (by minimum free energy) were compared; the optimal structure was retained unless it differed substantially (by visual comparison) from the Asteraceae model (Goertzen et al., 2003), in which case a suboptimal folding was used. Structure predictions for ITS2 also were validated against sequences in the ITS2 database (Schultz et al., 2006).

2.3. Character coding

The graphic output of Quikfold was converted into aligned nucleotide and pairwise interaction data for each stem-loop subregion (Figs. 1 and 2). Nucleotide interactions were coded numerically, using ‘1’ to indicate a complementary pairing and ‘0’ for a mismatch. Nucleotide and numerical data for multiple taxa were

Fig. 4. Single most-parsimonious phylogenetic tree derived from maximum parsimony analysis of numeric pairwise interaction data. Percent parsimony bootstrap support (above) and Bayesian posterior probability (below) values are given for each node. A phylogram for the same tree, with branch lengths, is depicted at right.
combined into a single matrix and aligned manually to maximize both structural and nucleotide similarity. Gaps that resulted from manual alignment (i.e., indels) were treated as missing data; they were encoded separately using a variation of simple indel coding (Simmons and Ochoterena, 2000). Indels were scored as present or absent, with an indel that spanned several consecutive nucleotides treated as a single character with states corresponding to the length of the indel. From the secondary structure data matrix, only numerical (pairwise interaction and indel) data were used for phylogeny reconstruction; the nucleotide identities of secondary structure interactions were retained for the purpose of ancestral state reconstruction (see below).

In order to evaluate the congruence of phylogeny estimation between structural and nucleotide characters, we analyzed linear

Fig. 5. Ancestral state reconstruction for subregion 1A of ITS1, using the tree topology and labeled nodes depicted in Fig. 4. Boxed nucleotides represent departures from the structure of the immediate ancestor (e.g., changes in the node F taxon relative to the node G taxon; the node J taxon was evaluated relative to the outgroup ancestral taxon). Deletion events are indicated with an ‘x’.
nucleotide sequences also. Alignment of nucleotide sequences was aided by the program POY (version 3.0.11; Wheeler, 1996; Wheeler et al. 2003), but the ultimate alignment was manual. Indels were not coded, due to high variability among sequences.

2.4. Phylogenetic analysis

Aligned structural and nucleotide matrices were analyzed under both maximum parsimony (MP) and Bayesian inference (BI) criteria. The parsimony analysis was conducted using PAUP* (version 4.0b10; Swoford, 2002). Partition-homogeneity/incongruence-length difference (ILD) tests (Farris et al., 1994) were implemented (heuristic search, 1000 replicates, maxtrees = 1000) after excluding constant and uninformative sites (Lee, 2001) using partitions among subregions for structural data and among ITS1, 5.8S, and ITS2 for nucleotide data, with an ILD exclusion threshold of $p < 0.01$. Data were evaluated for relative phylogenetic signal using the $g_1$ skewness statistic (Hillis and Huelsenbeck, 1992) by generating 100,000 random trees in PAUP* for each data subset. Phylogenetic trees were constructed using a full heuristic search.
in PAUP* (100 replicates of random stepwise addition, branch swapping by tree bisection and reconnection [TBR], maxtrees = 100,000), using Menyanthes–Nephrophyllidium for the outgroup (Lundberg and Bremer, 2003; Tippery et al., 2008). Support values for nodes were estimated using 1000 bootstrap replicates with the following options: heuristic search, one random stepwise addition per replicate, swapping by TBR, and maxtrees = 10,000.

Bayesian phylogenetic analysis was implemented using MrBayes (version 3.1.2; Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). Nucleotide data were partitioned with the following evolutionary models: SYM + C for ITS1, K80 + I for 5.8S, and GTR + I for ITS2, after model selection with Modeltest (version 3.4) under the AIC criterion (Posada and Crandall, 1998; Posada and Buckley, 2004; Posada, 2006). Structural data were analyzed using the ‘standard’ model with default parameters (Lewis, 2001). In each analysis, four independent runs of Markov Chain Monte Carlo (MCMC) were implemented with four heated chains each; trees were sampled every 1000th generation for 2,000,000 generations. The initial one-fourth of samples was discarded as burn-in.

2.5. Ancestral state reconstruction

Using the structural data matrix and the structural data maximum parsimony phylogeny (see Section 3), ancestral states were inferred using Mesquite (version 1.12; Maddison and Maddison, 2001) with likelihood ancestral states under the default model. The nucleotide identities of sites were reconstructed under parsimony using the ‘describe trees’ option with the ‘states for internal nodes’ output in PAUP* (Swofford, 2002); incompatible nucleotide pairings were amended to reflect plausible matches or mismatches, depending on the reconstructed structural state.

3. Results

Structural modeling of Menyanthaceae ITS sequences, using the algorithm of Zuker et al. (1999), predicted RNA secondary structures that conformed to the Asteraceae consensus model (Fig. 1; Goertzen et al., 2003). Several structural features were maintained in our analysis that had been identified by prior authors. Region 1C, which corresponds to a motif common among angiosperms (Liu and Schardl, 1994), and which Goertzen et al. (2003) found to be nearly invariant within Asteraceae, was highly conserved in Menyanthaceae also. Conserved portions of ITS2 included a 5′-UC opposite 3′-UU mismatch in region 2B, and a 5′-GGU site in region 2C, which were reported by Mai and Coleman (1997) in their survey across green plants. Structure prediction for ITS2 using the ITS2 database (Schultz et al., 2006) most often returned sequence and structure comparisons that were derived from species of Asteraceae, which was the plant family most abundantly represented in the database, out of taxa related to Menyanthaceae (Lundberg and Bremer, 2003; Tippery et al., 2008).

With respect to their aligned ITS1 and ITS2 nucleotide sequences (without considering structural data), Menyanthaceae taxa were 61–99% similar to each other in pairwise comparisons (p-distance). Species of Nymphoides were 84–98% similar to each other and 69–85% similar to species of Villarsia. In the 5.8S region, taxa were all >93% similar to each other. Aligned nucleotide data and structural data (pairwise interaction and indel) were submitted to TreeBASE (Study No. S2147). Character statistics for separate and combined data partitions are provided in Table 1. The following ILD p-values were obtained: structural data among all subregions (1A vs. 1B vs. 1C vs. 2A vs. 2B vs. 2C vs. 2D): 0.087; subsets of nucleotide data (ITS1 vs. 5.8S vs. ITS2): 0.995. In the partitioned Bayesian analysis of nucleotide data, the following parameters were estimated (with standard deviation in parentheses): alpha shape parameter for ITS1: 1.31 (0.28), proportion of invariant sites (pinvar) for 5.8S: 0.70 (0.03), pinvar for ITS2: 0.18 (0.04). Tree statistics for parsimony and Bayesian methods are given in Table 1.

Analysis of nucleotide data resolved the same topology that Tippery et al. (2008) reported in their study (Fig. 3), which differed only in having multiple accessions for some taxa. Most of the labeled internal nodes were resolved with high parsimony bootstrap (BS) and Bayesian posterior probability (PP) support (>80 BS/0.95

![Fig. 7. Ancestral state reconstruction for subregion 1C of ITS1, using the tree topology and labeled nodes depicted in Fig. 4. Boxed nucleotides represent departures from the structure of the immediate ancestor (e.g., changes in the node F taxon relative to the node G taxon; the node J taxon was evaluated relative to the outgroup ancestral taxon).](image-url)
PP), with the exception of the two nodes labeled B and D, which correspond, respectively, to the non-umbellate species of Nymphoides and the least well-resolved clade of Villarsia (also including Liparophyllum). The tree constructed using structural data (Fig. 4) had somewhat less resolution and lower support overall. The topologies of the nucleotide data and structural data trees were incongruent only with respect to Nymphoides crenata/N. peltata, and Villarsia reniformis/V. umbricola; however, topologies involving these taxa had only moderate support (<75% BS/0.80 PP) on the structural data tree (Figs. 3 and 4).

4. Discussion

Previous phylogenetic work on Menyanthaceae by Tippery et al. (2008) supported the monophyly of the genus Nymphoides but indicated that Villarsia and the monotypic genus Liparophyllum together comprise a paraphyletic grade. In their study, clades that were well supported on the ITS cladogram (cf. Fig. 3) were supported also by chloroplast molecular data and total combined data; however, the Nymphoides subclade of non-umbellate species (node B) and one of the three Villarsia clades (node D) were supported

![Ancestral state reconstruction for subregion 2A of ITS2, using the tree topology and labeled nodes depicted in Fig. 4. Boxed nucleotides represent departures from the structure of the immediate ancestor (e.g., changes in the node F taxon relative to the node G taxon; the node J taxon was evaluated relative to the outgroup ancestral taxon). Deletion events are indicated with an 'x'.](image-url)
only weakly by the ITS data and remained unresolved on the chloroplast data tree (Tippery et al., 2008). Data derived from predicted ITS secondary structures were used to evaluate relationships among these less closely related species (60–80% pairwise similarity between taxa), where phylogenetic analysis of ITS nucleotide sequences failed to yield significant topological support (Fig. 3), yet where structural similarity nonetheless could be determined through comparative analysis (Gutell et al., 2002; Goertzen et al., 2003).

The independent analysis of ITS pairwise interaction data produced a topology that was on the whole congruent with the cladogram constructed from nucleotide sequence data alone, and nodal support values were lower overall on the structural tree (Figs. 3 and 4). Most of the labeled internal nodes, which delimit major evolutionary groups in Menyanthaceae (Tippery et al., 2008), were supported by the structural data, except for node B, which was not recovered in the analysis (Fig. 4). Node B defines the clade of non-umbellate Nymphoides species, which received moderate support in a prior analysis of combined morphological and molecular data but was unresolved by chloroplast DNA data (Tippery et al., 2008). Subsequent analysis of additional Nymphoides taxa (N.P.T., unpublished data) failed to support the monophyly of non-umbellate species relative to the umbellate species, which accords with the result obtained from the analysis of ITS structural data.

Another internal node, representing a group of Villarsia species and Liparophyllum gunnii (node D), received moderate support in the structural data analysis (Fig. 4). The taxa were similarly unresolved by chloroplast data and supported moderately by combined data in a prior analysis (Tippery et al., 2008). Although nodal support was weak in both nucleotide data and structural data analyses, the clade was resolved consistently with relatively large branch lengths (Figs. 3 and 4). Structural characters that contributed to the resolution of node D (i.e., synapomorphic characters) were distributed evenly throughout the subregions of ITS1 and ITS2 (Figs. 5–11), indicating that several of the secondary structure subregions provided data to support the monophyly of the node. Nucleotide changes that did not alter the predicted structures had no additional cost in our analysis, but substitutions that disrupted secondary structure were penalized. Data that resolved node D thus represent rare changes in a highly conserved structural region, not unlike the phylogenetic data that often are used in higher-level taxonomic comparisons (Coleman, 2003).

Our analysis of ITS structural data differed substantially from other methods that have modeled the secondary structure of ribosomal RNA genes in a phylogenetic context (Wheeler and Honeycutt, 1988; Steele et al., 1991; Dixon and Hillis, 1993; Kjer, 1995; Schöniger and von Haeseler, 1999; Gutell et al., 2002). Rather than generating a consensus structural model for all taxa, we allowed individual sites to be paired or unpaired and structural elements in different taxa to be composed of non-homologous nucleotides. We thus decoupled nucleotide and structural data from the one-to-one relationship under which they usually are analyzed. Although predicted ITS secondary structures depend explicitly on underlying nucleotide sequences, the two data types could differ where either changes in nucleotide sequence have no effect on structure or homologous strings of nucleotides compose different structural elements in different taxa. The structural data in our study had strong phylogenetic signal (measured by the g1 skewness statistic; Table 1; Hillis and Huelsenbeck, 1992) and recovered nearly the same tree topology as the nucleotide data (Figs. 3 and 4).

The ITS structural data we encoded arguably represent an independent set of data from the nucleotide data. Although ITS secondary structure clearly depends upon the component sequence of nucleotides, there are separate evolutionary and selective pressures that operate at each level. Strings of one or more nucleotides are altered by single base pair changes or indel events, and at some frequency such mutational changes are incorporated into DNA sequences. Certain mutations engender a downstream change in secondary structure. Whether structural changes are retained or purged from populations depends in part on selective forces that result from functional constraints on the RNA molecule. With respect to ITS, some regions are conserved for sequence (Liu and Schardl, 1994; Mai and Coleman, 1997), whereas others apparently reflect selection on secondary structure irrespective of sequence (van Nues et al., 1994, 1995; Joseph et al., 1999; Michot et al.,
In the former case, structure is tied directly to nucleotide sequence, and the alteration of a single base pair could disrupt a conserved structure. In the latter, however, of which several examples have been uncovered, strings of nucleotides on one RNA strand can pair with alternate strings on the complementary strand and retain the same overall structure. When ITS2 structures were compared across the most divergent eukaryote taxa, for example, a consistent secondary structure emerged that nonetheless reflected a vast amount of underlying nucleotide variation (Schultz et al., 2005). The complexity of interactions between single base pairs and the structures they encode, then, often would exceed the amount of data contained in a simple nucleotide or structure alignment. Our analysis thus attempted to account for structure conservation without constraining positionally homologous nucleotides to produce the same structure in every taxon. Because of the different stochastic and selective factors that affect nucleotide sequence vs. secondary structure, the two data sets used in our analysis could be considered independent from each other, in which case they could be combined into a single data matrix. In order to evaluate the relative contributions of nucleotide and structural data to phylogeny estimation, we analyzed a matrix of combined nucleotide and structural data (not shown), which resulted in a topology

Fig. 10. Ancestral state reconstruction for subregion 2C of ITS2, using the tree topology and labeled nodes depicted in Fig. 4. Boxed nucleotides represent departures from the structure of the immediate ancestor (e.g., changes in the node F taxon relative to the node G taxon; the node J taxon was evaluated relative to the outgroup ancestral taxon). Deletion events are indicated with an ‘x’.
that was identical to the nucleotide data tree (Fig. 3), in which par-
simony bootstrap support for nodes B and D increased from 58% to
78% and from 77% to 95%, respectively. The application of com-
bined nucleotide and structural data may warrant further consid-
eration, after determining to what extent the two data types are
interdependent.

In addition to phylogeny estimation, the ITS structural data
were useful for the purpose of reconstructing putative character
states for hypothesized ancestral taxa, using the Menyanthaceae
tree topology obtained from structural data (Fig. 4). Ancestral char-
acter states for structural RNA have been reconstructed in a phylo-
genetic context previously by Hickson et al. (1996), who analyzed
compensatory base pair changes in stem regions of rDNA (see also
Sluiman et al., 2008); however, in our study, where ITS nucleotide
sequences were highly divergent even for closely related taxa, we
focused on reconstructing only changes in structure. In an example
from subregion 1A of ITS1, several motifs were identified that were
conserved in all Menyanthaceae taxa, consisting of both structural
and nucleotide conservation (Fig. 5). Furthermore, a number of
structural changes persisted through several descendent nodes or
were synapomorphic for taxa belonging to a particular clade. For
example, the two unpaired sites highlighted for node F are indica-
tive of the descendant species Villarsia capensis and V. manningiana,
and the unpaired nucleotide highlighted for node E represents a
shared ancestral state for taxa descended from nodes C and D. Dif-
ferences in structure among ancestral taxa often were brought
about by disrupting or reestablishing pairwise complementarity,
or by inserting or deleting one or a few nucleotides. In the 1A sub-
region, there were no obvious shifts among strings of paired nucle-
otides (i.e., pairing between non-homologous nucleotides in
different taxa), although such changes would have had no effect
on the encoded numerical data if they preserved the secondary
structure. Structural changes among ancestral taxa were highly
conserved and seldom reversed, providing a strong phylogenetic
signal with which to define descendent clades.

Deriving secondary structure characters from mathematically
predicted models depends heavily on accurate sequencing of the
ITS regions and reliable secondary structure prediction. RNA struc-
ture predictions are extremely sensitive to single nucleotide differ-
ences, which may result in the disruption of site pairing, or more
seriously, in the shift of paired nucleotides along the stem (Kjer,
1995; Hickson et al., 1996; Mai and Coleman, 1997). Consequently,
the sequences used in our analysis were meticulously evaluated by
eye for signal quality and accurate nucleotide assignment in order
to avoid erroneous structural predictions. We acknowledge that
mathematical algorithms also are imperfect predictors of structure
(Mathews et al., 1999), and to date no crystal structure of either ITS
region has been resolved, against which predicted structures could
be evaluated. Phylogenetic analysis of predicted RNA structural
features, however, could provide valuable feedback for thermody-
namic modeling and help generate more accurate structure predic-
tions in the future. Furthermore, examining ITS secondary
structures in a phylogenetic context should encourage additional
research into their functional significance.

The described method could be used in combination with avail-
able software packages that align both nucleotide sequence and
predicted secondary structure. The widely implemented ‘Vienna
string’ notation (Hofacker et al., 1994) could be converted into a
structural data matrix by replacing each paired site (indicated with
parentheses, ‘(‘ or ‘)’) with a ‘1’ and each unpaired site (noted by a
period, '.') with a ‘0’. In addition, output from the structure predic-
tion module of the ITS2 database (Schultz et al., 2006), for example,
generates indels among compared taxa, which could be encoded
and analyzed similarly to the method we have described. If more
widely implemented and more thoroughly refined, the method
should become a useful tool for extracting additional phylogenetic
signal from the often utilized but poorly understood internal tran-
scribed spacer.

Acknowledgments

The authors are indebted to K. Kjer, C. Simon, and two anony-
mous reviewers for their helpful comments on earlier drafts of
the method and manuscript.
References


