



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

ARTICLE INFO

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

ABSTRACT

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 (Zuckerandl 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, 1999; 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

* Corresponding author. Fax: +1 860 486 6364.

E-mail address: nicholas.tippery@uconn.edu (N.P. Tippery).

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 (Biloud 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

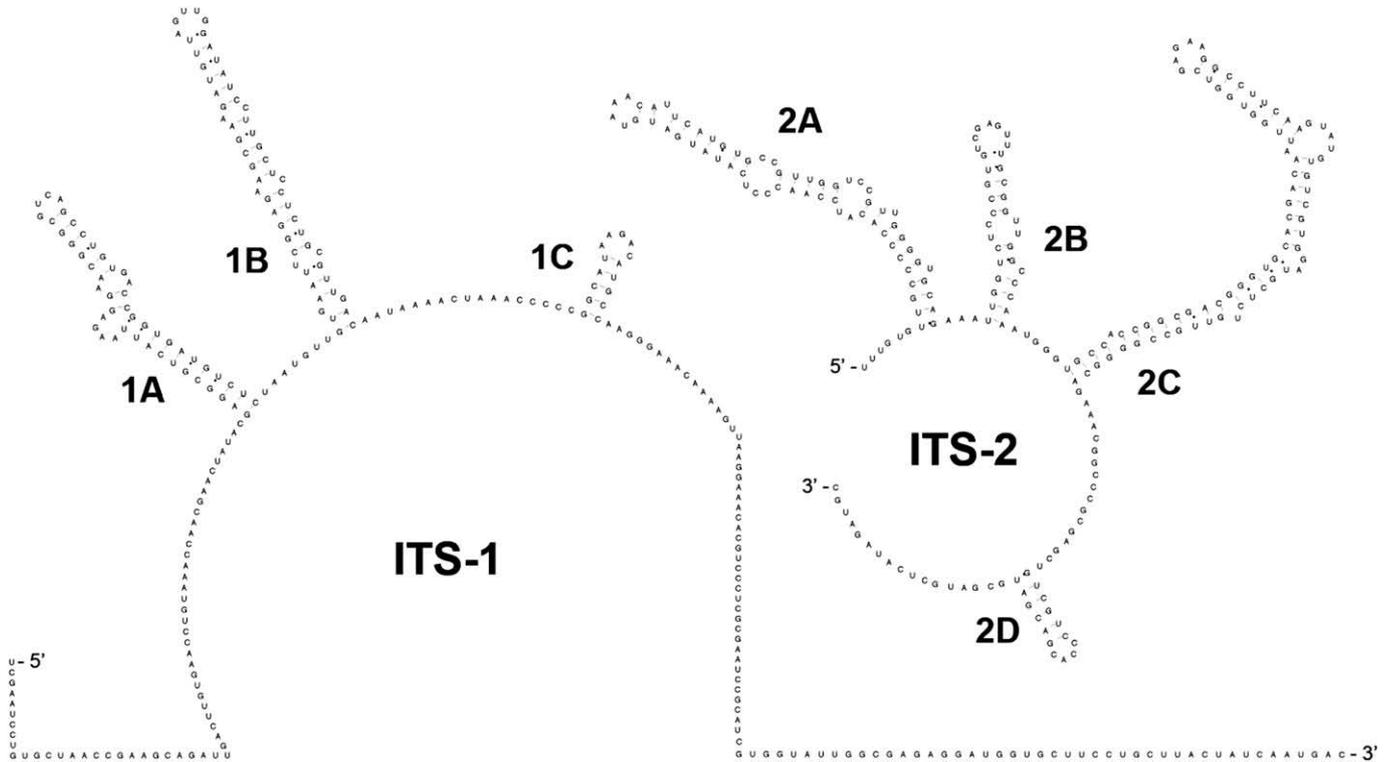


Fig. 1. Predicted secondary structure for the *Nymphoides cordata* ITS1 and ITS2 regions (the intervening 5.8S rRNA has been omitted). Stem subregions correspond to conserved structures identified by Goertzen et al. (2003) for Asteraceae. Illustration produced using the program XRNA (B. Weiser and H.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

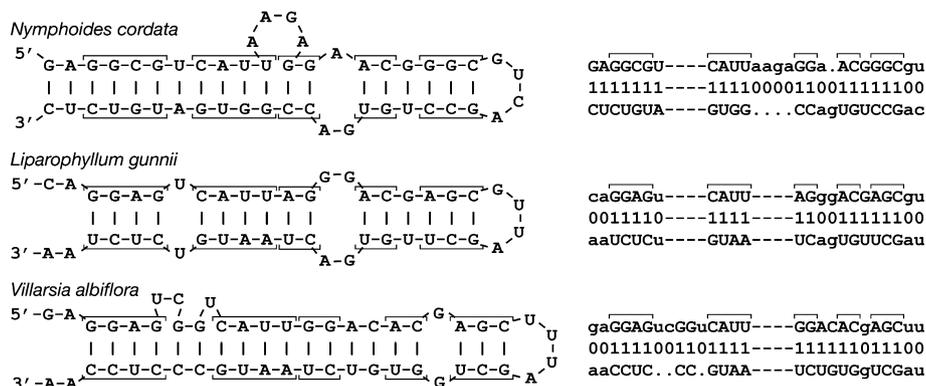


Fig. 2. Summary of the character encoding method. Example predicted secondary structures for the 1A subregion of ITS1 are given at left for three taxa. Conserved stem portions (i.e., found in all example taxa) are bracketed in the predicted structures and in the encoded and aligned data at right. For 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 (Höschmann 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; Guttell 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
Statistics for phylogenetic trees generated from nucleotide sequence and structural data under maximum parsimony (MP) and Bayesian inference (BI) methods

Data type	# Characters	# Parsimony informative (%)	g_1	# Trees (MP)	Tree length (MP)	CI (MP)	RI (MP)	CI _{exc} (MP)	lnL (BI)
ITS1	353	151 (43)	-0.65						
5.8S	168	11 (07)	-0.62						
ITS2	290	129 (44)	-0.84						
Total nucleotide	811	291 (36)	-0.75	6	742	0.74	0.85	0.69	-4672
1A	38	18 (47)	-0.35						
1B	47	22 (47)	-0.56						
1C	8	2 (25)	-0.84						
2A	52	22 (42)	-0.65						
2B	22	5 (23)	-0.55						
2C	62	26 (42)	-0.71						
2D	13	4 (31)	-2.45						
Total structural	243	99 (41)	-0.72	1	277	0.58	0.75	0.52	-1146

CI, consistency index; RI, retention index; CI_{exc}, CI excluding uninformative characters; lnL, natural log likelihood (harmonic mean).

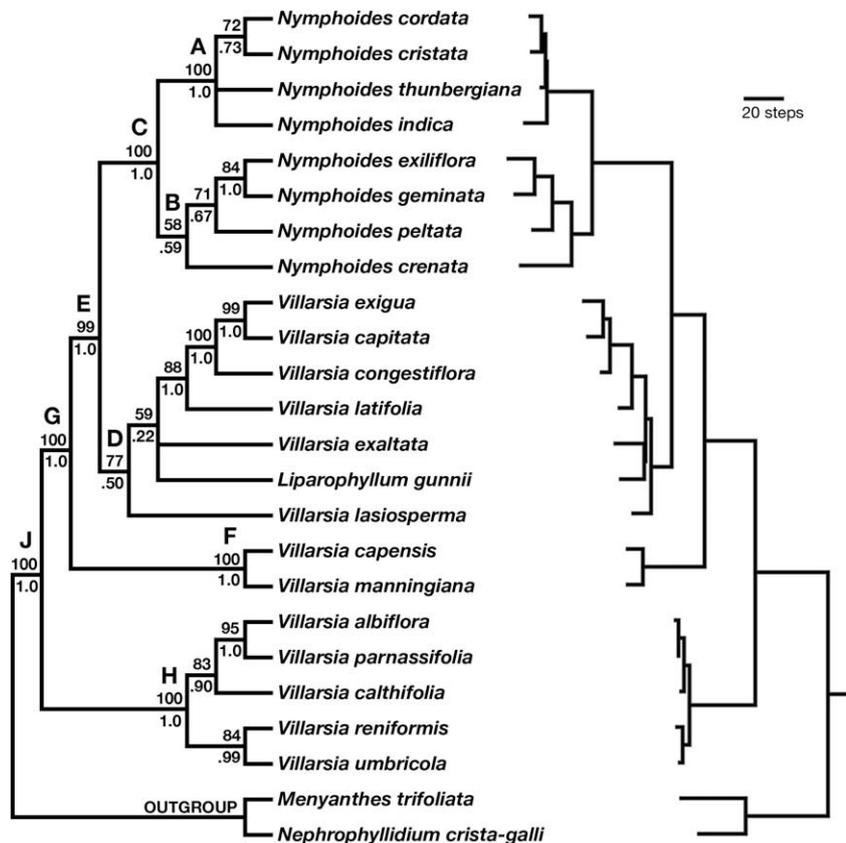


Fig. 3. Strict consensus phylogenetic tree (left) derived from maximum parsimony analysis of ITS nucleotide sequences (including the 5.8S rDNA gene). Percent parsimony bootstrap support (above) and Bayesian posterior probability (below) values are given for each node. A phylogram for one of six most-parsimonious 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 *Nephrophyllidium*), 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 DINAMelt 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

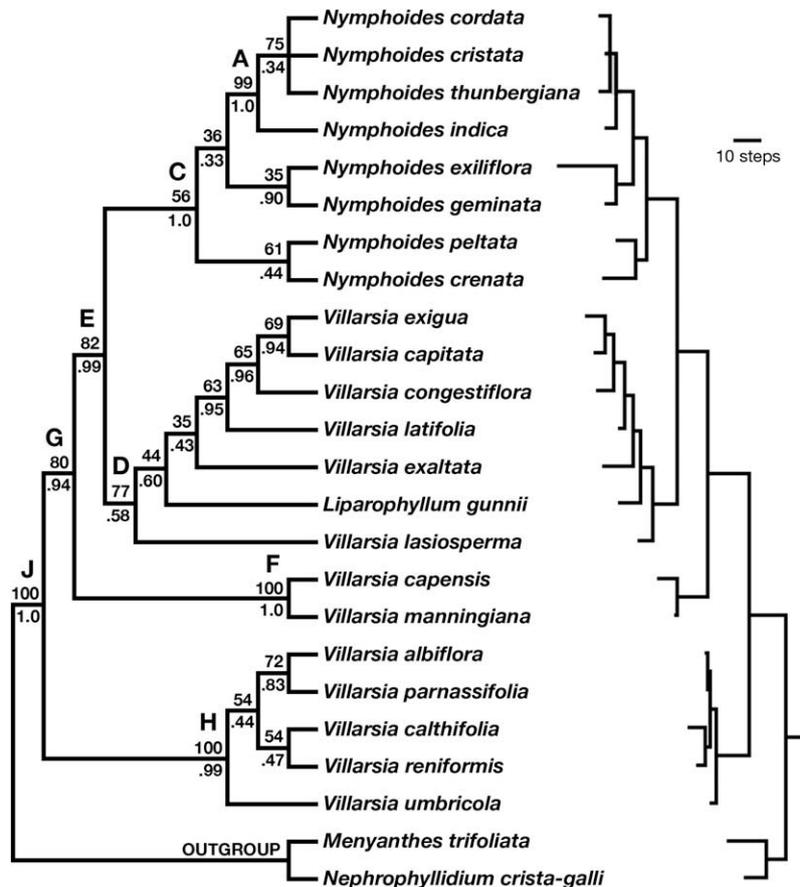


Fig. 4. Single most-parsimonious phylogenetic tree derived from maximum parsimony analysis of numeric pairwise interaction data. Percent parsimony bootstrap support (above) and Bayesian posterior probability (below) values are given for each node. A phylogram for the same tree, with branch lengths, is 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

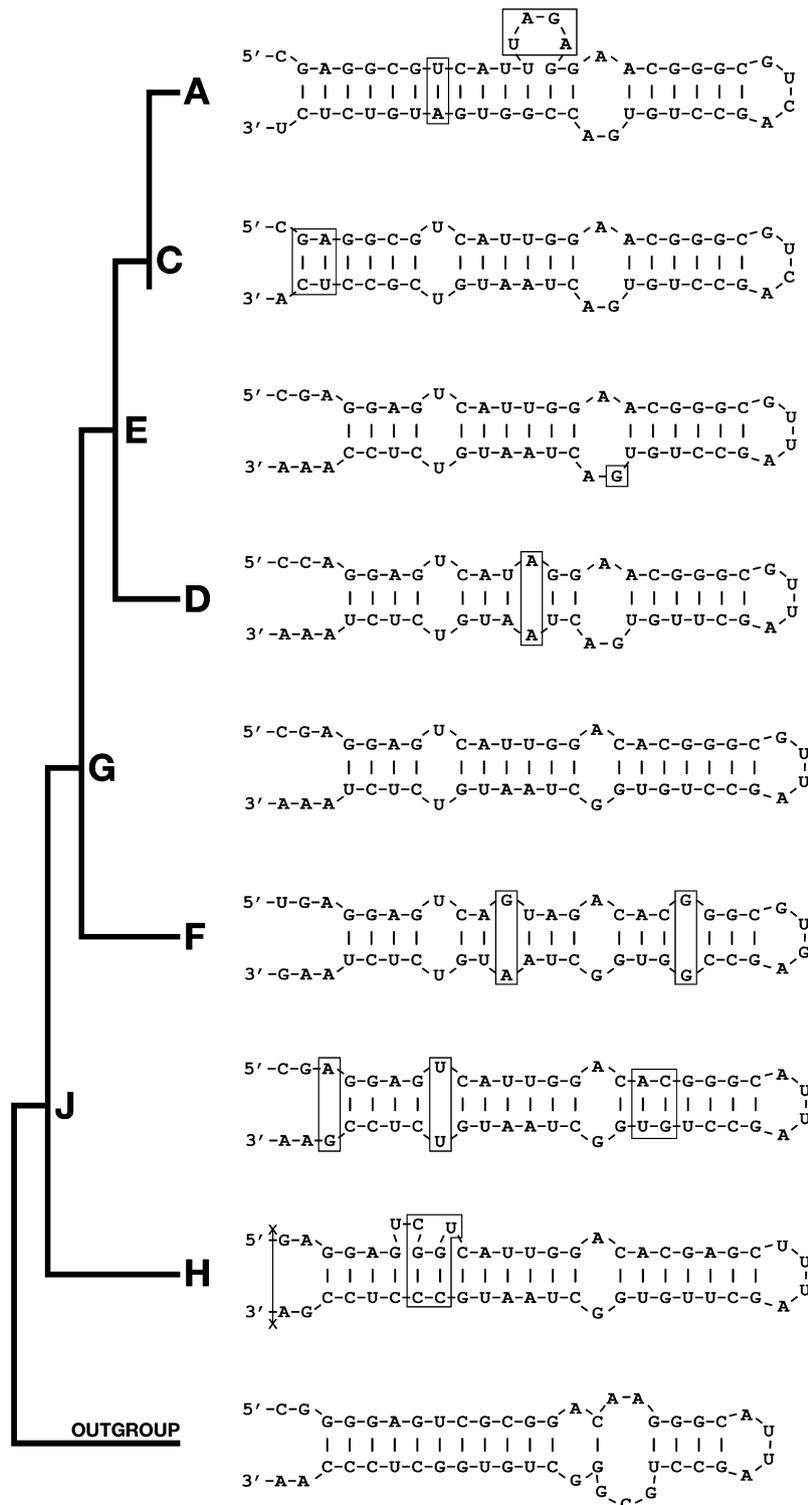


Fig. 5. Ancestral state reconstruction for subregion 1A of ITS1, using the tree topology and labeled nodes depicted in Fig. 4. Boxed nucleotides represent departures from the structure of the immediate ancestor (e.g., changes in the node F taxon relative to the node G taxon; the node J taxon was evaluated 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; Swofford, 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

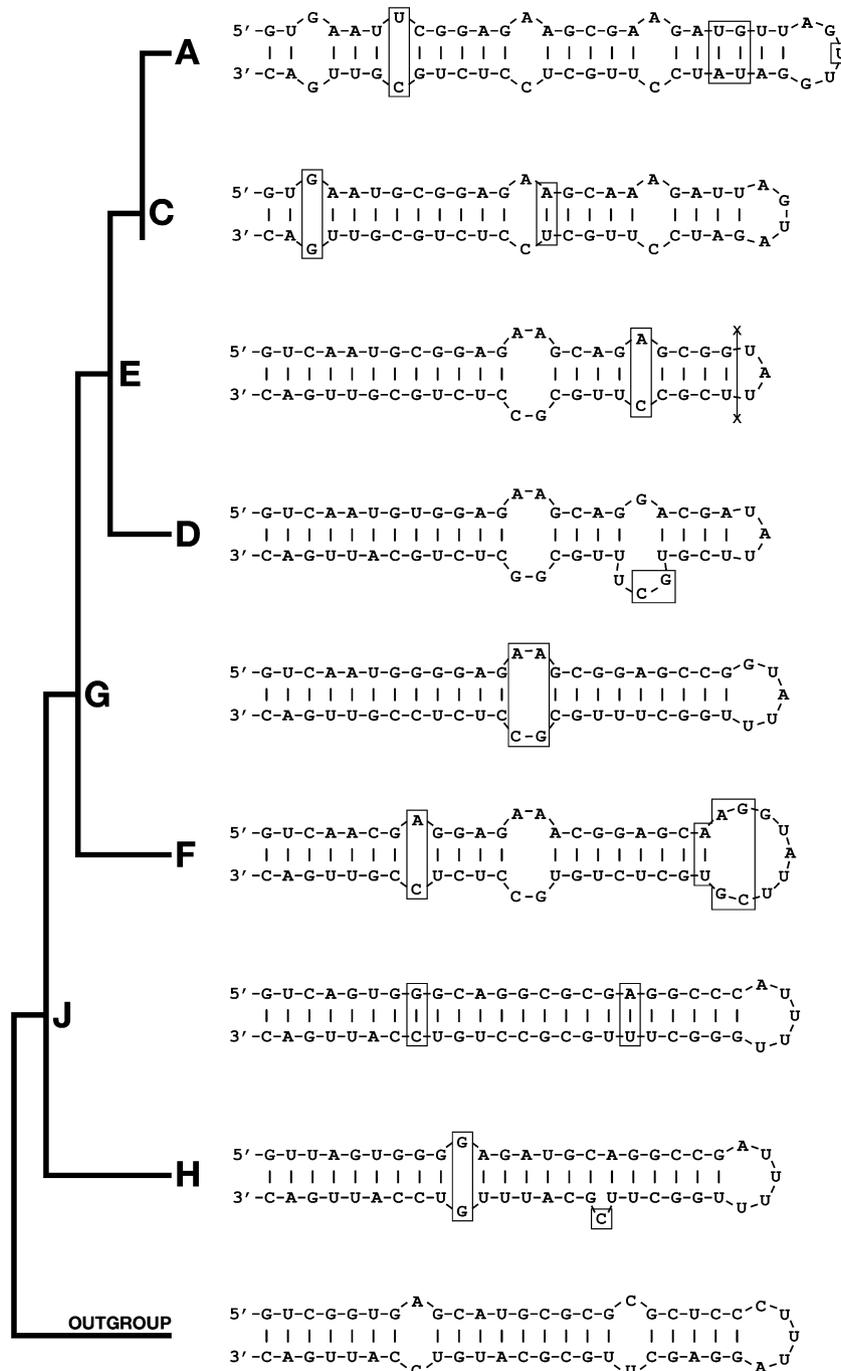


Fig. 6. Ancestral state reconstruction for subregion 1B of ITS1, using the tree topology and labeled nodes depicted in Fig. 4. Boxed nucleotides represent departures from the structure of the immediate ancestor (e.g., changes in the node F taxon relative to the node G taxon; the node J taxon was evaluated relative to the outgroup ancestral taxon). Deletion events are indicated with an 'X'.

in PAUP* (100 replicates of random stepwise addition, branch swapping by tree bisection and reconnection [TBR], maxtrees = 100,000), using *Menyanthes–Neprophyllidium* 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 + Γ 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'-UC or -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

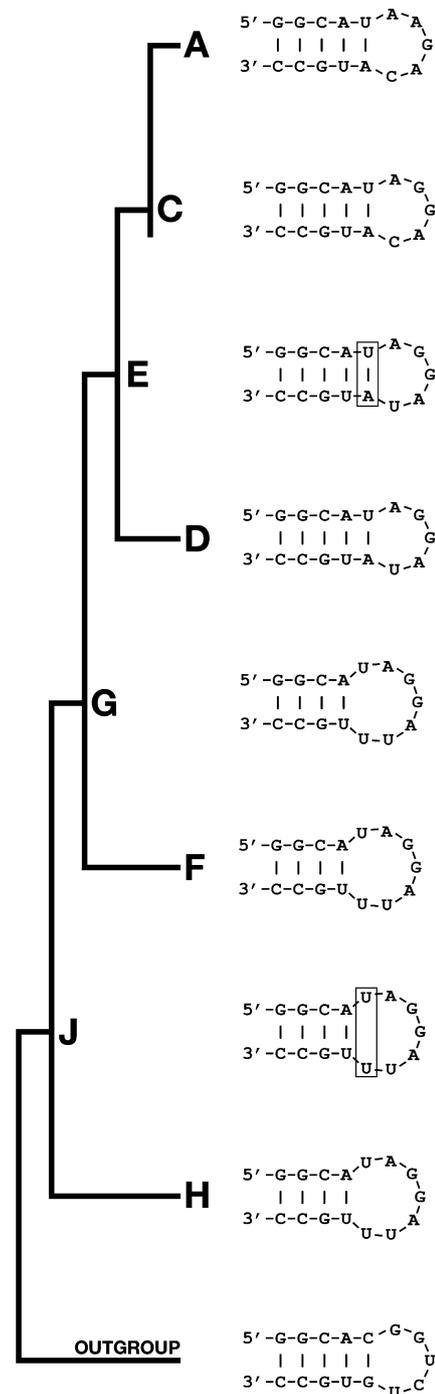


Fig. 7. Ancestral state reconstruction for subregion 1C of ITS1, using the tree topology and labeled nodes depicted in Fig. 4. Boxed nucleotides represent departures from the structure of the immediate ancestor (e.g., changes in the node F taxon relative to the node G taxon; the node J taxon was evaluated relative to the outgroup ancestral taxon).

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

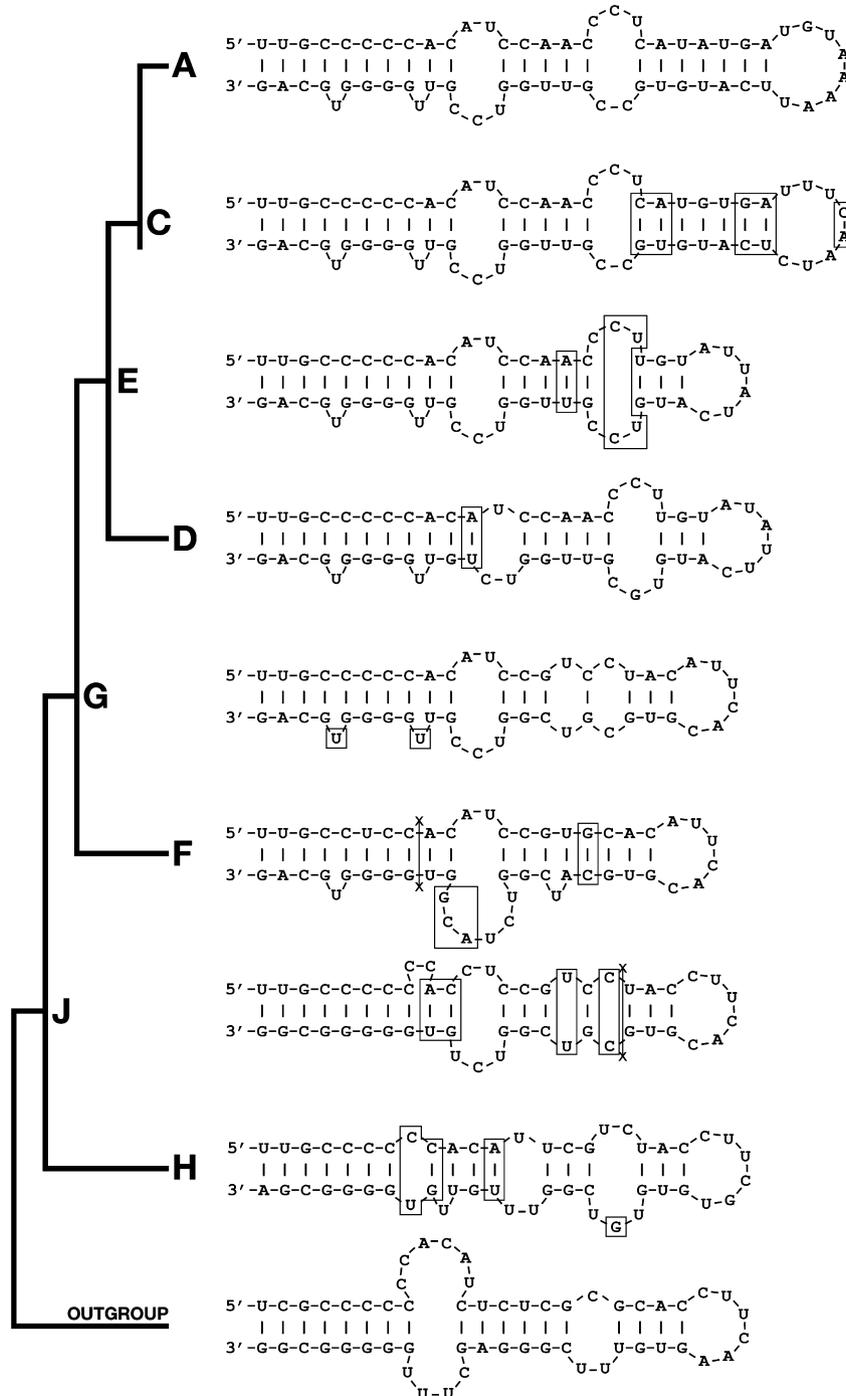


Fig. 8. Ancestral state reconstruction for subregion 2A of ITS2, using the tree topology and labeled nodes depicted in Fig. 4. Boxed nucleotides represent departures from the structure of the immediate ancestor (e.g., changes in the node F taxon relative to the node G taxon; the node J taxon was evaluated relative to the outgroup ancestral taxon). Deletion events are indicated with an 'X'.

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 g_1 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

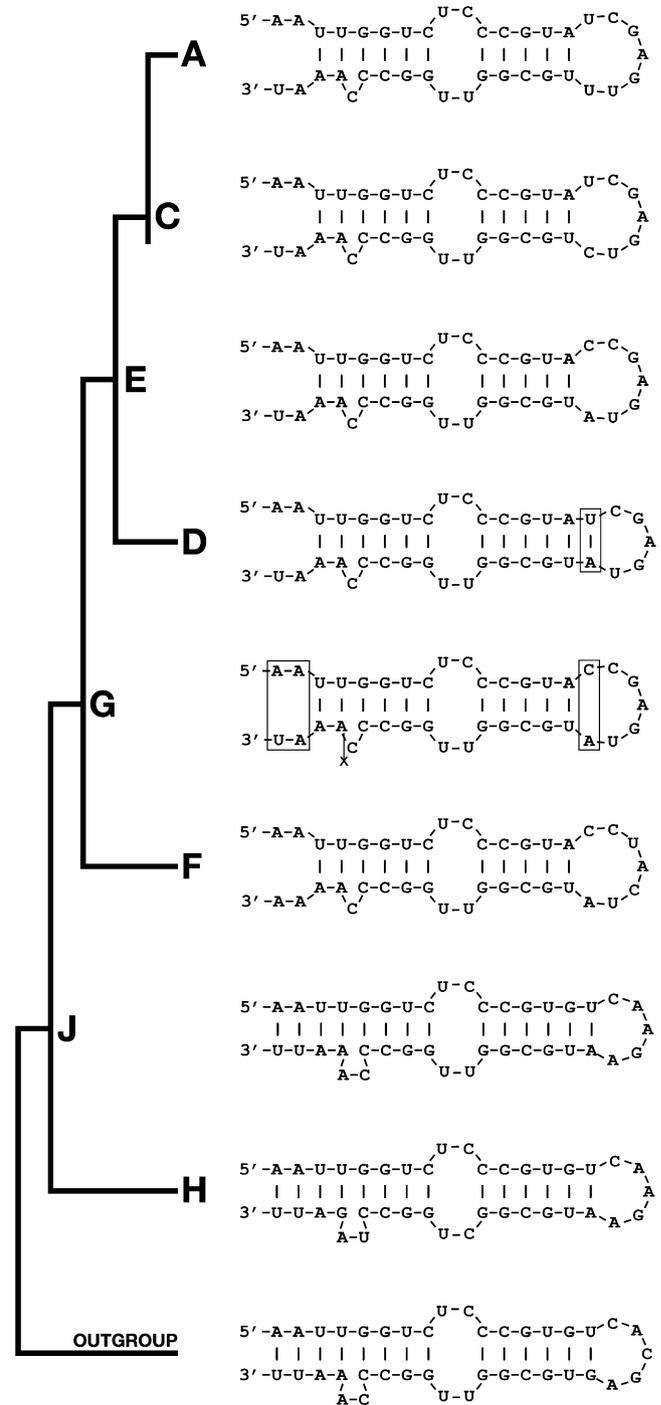


Fig. 9. Ancestral state reconstruction for subregion 2B of ITS2, using the tree topology and labeled nodes depicted in Fig. 4. Boxed nucleotides represent departures from the structure of the immediate ancestor (e.g., changes in the node F taxon relative to the node G taxon; the node J taxon was evaluated relative to the outgroup ancestral taxon). Events are indicated with an 'x'.

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.,

1999). 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

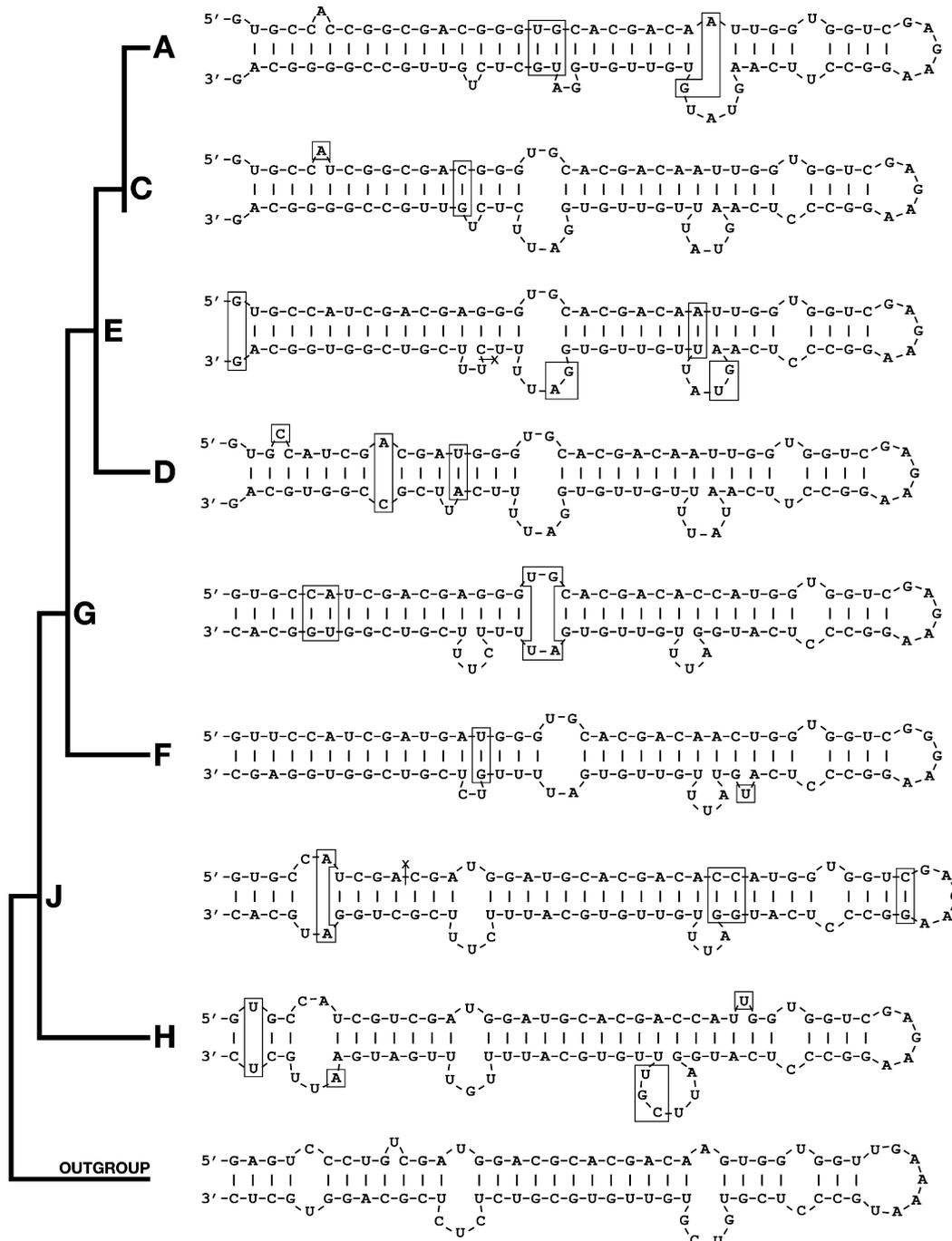


Fig. 10. Ancestral state reconstruction for subregion 2C of ITS2, using the tree topology and labeled nodes depicted in Fig. 4. Boxed nucleotides represent departures from the structure of the immediate ancestor (e.g., changes in the node F taxon relative to the node G taxon; the node J taxon was evaluated relative to the outgroup ancestral taxon). Deletion events are indicated with an 'X'.

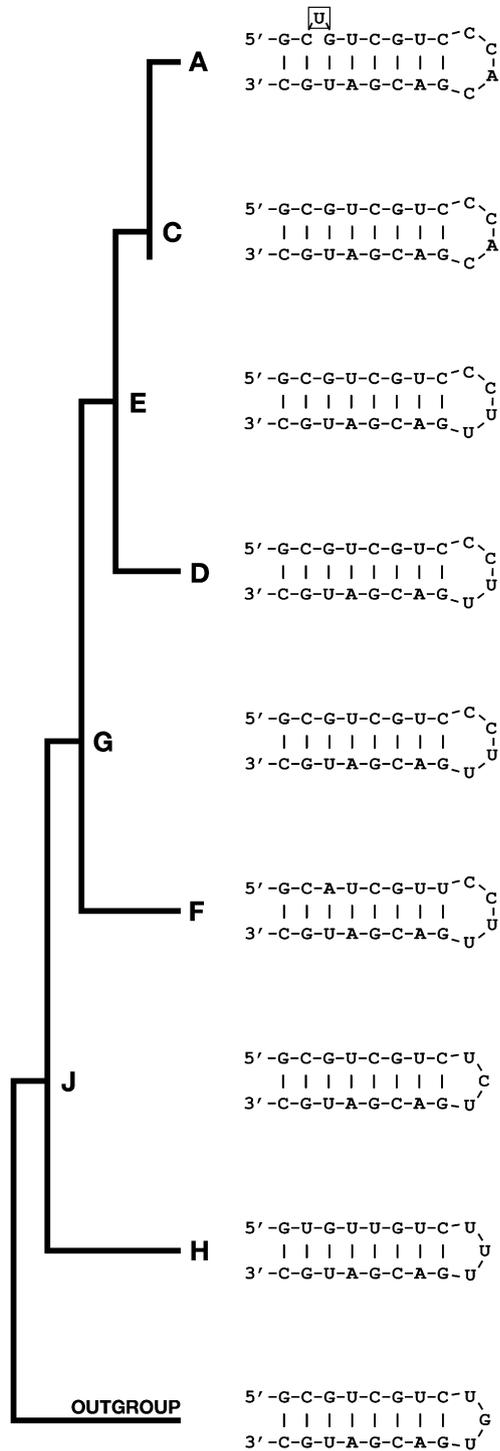


Fig. 11. Ancestral state reconstruction for subregion 2D of ITS2, using the tree topology and labeled nodes depicted in Fig. 4. Boxed nucleotides represent departures from the structure of the immediate ancestor (e.g., changes in the node F taxon relative to the node G taxon; the node J taxon was evaluated relative to the outgroup ancestral taxon).

that was identical to the nucleotide data tree (Fig. 3), in which parsimony bootstrap support for nodes B and D increased from 58% to 78% and from 77% to 95%, respectively. The application of combined nucleotide and structural data may warrant further consideration, 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 character states for structural RNA have been reconstructed in a phylogenetic 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 indicative of the descendent 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. Differences 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 nucleotides (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 structure predictions are extremely sensitive to single nucleotide differences, 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 thermodynamic modeling and help generate more accurate structure predictions 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 available 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 prediction 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 transcribed spacer.

Acknowledgments

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

References

- Billoud, B., Guerrucci, M.-A., Masselot, M., Deutsch, J.S., 2000. Cirripede phylogeny using a novel approach: molecular morphometrics. *Mol. Biol. Evol.* 17, 1435–1445.
- Buckley, T.R., Simon, C., Chambers, G.K., 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. *Syst. Biol.* 50, 67–86.
- Caetano-Anollés, G., 2002. Evolved RNA secondary structure and the rooting of the universal tree of life. *J. Mol. Evol.* 54, 333–345.
- Chen, C.A., Chang, C.-C., Wei, N.V., Chen, C.-H., Lein, Y.-T., Lin, H.-E., Dai, C.-F., Wallace, C.C., 2004. Secondary structure and phylogenetic utility of the ribosomal internal transcribed spacer 2 (ITS2) in scleractinian corals. *Zool. Stud.* 43, 759–771.
- Coleman, A.W., 2003. ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends Genet.* 19, 370–375.
- Coleman, A.W., Preparata, R.M., Mehrotra, B., Mai, J.C., 1998. Derivation of the secondary structure of the ITS-1 transcript in Volvocales and its taxonomic correlations. *Protist* 149, 135–146.
- Côté, C.A., Greer, C.L., Peculis, B.A., 2002. Dynamic conformational model for the role of ITS2 in pre-rRNA processing in yeast. *RNA* 8, 786–797.
- Dixon, M.T., Hillis, D.M., 1993. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. *Mol. Biol. Evol.* 10, 256–267.
- Farris, J.S., Källersjö, M., Kluge, A.G., Bult, C., 1994. Constructing a significance test for incongruence. *Syst. Biol.* 44, 570–572.
- Fougère-Danezan, M., Maumont, S., Bruneau, A., 2007. Relationships among resin-producing Detarieae s.l. (Leguminosae) as inferred by molecular data. *Syst. Bot.* 32, 748–761.
- Goertzen, L.R., Cannone, J.J., Gutell, R.R., Jansen, R.K., 2003. ITS secondary structure derived from comparative analysis: implications for sequence alignment and phylogeny of the Asteraceae. *Mol. Phylogenet. Evol.* 29, 216–234.
- Gottschling, M., Hilger, H.H., Wolf, M., Diane, N., 2001. Secondary structure of the ITS1 transcript and its application in a reconstruction of the phylogeny of Boraginales. *Plant Biol.* 3, 629–636.
- Gutell, R.R., Lee, J.C., Cannone, J.J., 2002. The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.* 12, 301–310.
- Hershkovitz, M.A., Lewis, L.A., 1996. Deep-level diagnostic value of the rDNA-ITS region. *Mol. Biol. Evol.* 13, 1276–1295.
- Hershkovitz, M.A., Zimmer, E.A., 1996. Conservation patterns in angiosperm rDNA ITS2 sequences. *Nucleic Acids Res.* 24, 2857–2867.
- Hickson, R.E., Simon, C., Cooper, A., Spicer, G.S., Sullivan, J., Penny, D., 1996. Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. *Mol. Biol. Evol.* 13, 150–169.
- Hillis, D.M., Huelsenbeck, J.P., 1992. Signal, noise, and reliability in molecular phylogenetic analyses. *J. Hered.* 83, 189–195.
- Höchsmann, M., Toller, T., Giegerich, R., Kurtz, S., 2003. Local similarity in RNA secondary structures. *Proc. IEEE Comput. Soc. Bioinform. Conf.* 2, 159–168.
- Höchsmann, M., Voss, B., Giegerich, R., 2004. Pure multiple RNA secondary structure alignments: A progressive profile approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1, 53–62.
- Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., Schuster, P., 1994. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* 125, 167–188.
- Hofacker, I.L., Fekete, M., Stadler, P.F., 2002. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* 319, 1059–1066.
- Huelsenbeck, J.P., Ronquist, F., 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755.
- Joseph, N., Krauskopf, E., Vera, M.I., Michot, B., 1999. Ribosomal internal transcribed spacer 2 (ITS2) exhibits a common core of secondary structure in vertebrates and yeast. *Nucleic Acids Res.* 27, 4533–4540.
- Kjer, K.M., 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. *Mol. Phylogenet. Evol.* 4, 314–330.
- Kjer, K.M., 2004. Aligned 18S and insect phylogeny. *Syst. Biol.* 53, 506–514.
- Krüger, D., Gargas, A., 2008. Secondary structure of ITS2 rRNA provides taxonomic characters for systematic studies—a case in Lycoperdaceae (Basidiomycota). *Mycol. Res.* 112, 316–330.
- Lee, M.S.Y., 2001. Uninformative characters and apparent conflict between molecules and morphology. *Mol. Biol. Evol.* 18, 676–680.
- Lewis, P.O., 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* 50, 913–925.
- Liu, J.-S., Schardl, C.L., 1994. A conserved sequence in internal transcribed spacer 1 of plant nuclear rRNA genes. *Plant Mol. Biol.* 26, 775–778.
- Lundberg, J., Bremer, K., 2003. A phylogenetic study of the order Asterales using one morphological and three molecular data sets. *Int. J. Plant Sci.* 164, 553–578.
- Maddison, W.P., Maddison, D.R., 2001. Mesquite: a modular system for evolutionary analysis. Available from: <<http://mesquite.biosci.arizona.edu/mesquite/mesquite.html>>.
- Mai, J.C., Coleman, A.W., 1997. The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. *J. Mol. Evol.* 44, 258–271.
- Markham, N.R., Zuker, M., 2005. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.* 33, W577–W581.
- Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H., 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288, 911–940.
- Michot, B., Joseph, N., Mazan, S., Bachelier, J.P., 1999. Evolutionarily conserved structural features in the ITS2 of mammalian pre-rRNAs and potential interactions with the snoRNA U8 detected by comparative analysis of new mouse sequences. *Nucleic Acids Res.* 27, 2271–2282.
- Posada, D., 2006. ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Res.* 34 (Web Server issue), W700–W703.
- Posada, D., Buckley, T.R., 2004. Model selection and model averaging in phylogenetics: advantages of the AIC and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* 53, 793–808.
- Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14, 817–818.
- Powell, J.R., Moriyama, E.N., 1997. Evolution of codon usage bias in *Drosophila*. *Proc. Natl. Acad. Sci. USA* 94, 7784–7790.
- Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574.
- Schöniger, M., von Haeseler, A., 1994. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.* 3, 240–247.
- Schöniger, M., von Haeseler, A., 1999. Toward assigning helical regions in alignments of ribosomal RNA and testing the appropriateness of evolutionary models. *J. Mol. Evol.* 49, 691–698.
- Schultz, J., Maisel, S., Gerlach, D., Müller, T., Wolf, M., 2005. A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *RNA* 11, 361–364.
- Schultz, J., Müller, T., Achtziger, M., Seibel, P.N., Dandekar, T., Wolf, M., 2006. The internal transcribed spacer 2 database—a web server for (not only) low level phylogenetic analyses. *Nucleic Acids Res.* 34, W704–W707.
- Seibel, P.N., Müller, T., Dandekar, T., Schultz, J., Wolf, M., 2006. 4SALE: a tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics* 7, 498.
- Siebert, S., Backofen, R., 2005. MARNAs: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics* 21, 3352–3359.
- Simmons, M.P., Ochoterena, H., 2000. Gaps as characters in sequence-based phylogenetic analyses. *Syst. Biol.* 49, 369–381.
- Sluiman, H.J., Guihal, C., Mudimu, O., 2008. Assessing phylogenetic affinities and species delimitations in Klebsormidiales (Streptophyta): nuclear-encoded rDNA phylogenies and ITS secondary structure models in *Klebsormidium*, *Hormidiella*, and *Entransia*. *J. Phycol.* 44, 183–195.
- Steele, K.P., Holsinger, K.E., Jansen, R.K., Taylor, D.W., 1991. Assessing the reliability of 5S rRNA sequence data for phylogenetic analysis in green plants. *Mol. Biol. Evol.* 8, 240–248.
- Swofford, D.L., 2002. PAUP*. Phylogenetic analysis using parsimony (* and other methods), version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.
- Tippery, N.P., Les, D.H., Padgett, D.J., Jacobs, S.W.L., 2008. Generic circumscription in Menyanthaceae: a phylogenetic evaluation. *Syst. Bot.* 33, 598–612.
- van Nues, R.W., Rientjes, J.M.J., van der Sande, C.A.F.M., Zerp, S.F., Sluiter, C., Venema, J., Planta, R.J., Raué, H.A., 1994. Separate structural elements within internal transcribed spacer 1 of *Saccharomyces cerevisiae* precursor ribosomal RNA direct the formation of 17S and 26S rRNA. *Nucleic Acids Res.* 22, 912–919.
- van Nues, R.W., Rientjes, J.M.J., Morré, S.A., Mollee, E., Planta, R.J., Venema, J., Raué, H.A., 1995. Evolutionarily conserved structural elements are critical for processing of internal transcribed spacer 2 from *Saccharomyces cerevisiae* precursor ribosomal RNA. *J. Mol. Biol.* 250, 24–36.
- Venema, J., Tollervy, D., 1999. Ribosome synthesis in *Saccharomyces cerevisiae*. *Annu. Rev. Genet.* 33, 261–311.
- Wang, S., Bao, Z., Li, N., Zhang, L., Hu, J., 2007. Analysis of the secondary structure of ITS1 in Pectinidae: implications for phylogenetic reconstruction and structural evolution. *Mar. Biotechnol.* 9, 231–242.
- Wheeler, W.C., 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? *Cladistics* 12, 1–9.
- Wheeler, W.C., Honeycutt, R.L., 1988. Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implications. *Mol. Biol. Evol.* 5, 90–96.
- Wheeler, W.C., Gladstein, D., De Laet, J., 2003. POY. Phylogenetic reconstruction via optimization of DNA and other data, version 3.0.11. Available from: <<http://research.amnh.org/scicomp/projects/poy.php>>.
- Wolf, M., Achtziger, M., Schultz, J., Dandekar, T., Müller, T., 2005. Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. *RNA* 11, 1616–1623.
- Zuckerandl, E., Pauling, L., 1965. Evolutionary divergence and convergence in proteins. In: Bryson, V., Vogel, H.J. (Eds.), *Evolving Genes and Proteins*. Academic Press, New York.
- Zuker, M., 1989. On finding all suboptimal foldings of an RNA molecule. *Science* 244, 48–52.
- Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406–3415.
- Zuker, A.M., Mathews, B.D.H., Turner, C.D.H., 1999. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: Barciszewski, J., Clark, B.F.C. (Eds.), *RNA Biochemistry and Biotechnology*, Number 70 in NATO Science Partnership Sub-Series: 3: High Technology. Kluwer Academic Publishers, Dordrecht.