Molecular Phylogenetics and Evolution 50 (2009) 364–375

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

# Comparative study of the evolution of nuclear ribosomal spacers incorporating secondary structure analyzes within Dodonaeoideae, Hippocastanoideae and Xanthoceroideae (Sapindaceae)

Mark G. Harrington [a,*], Ed Biffin [b], Paul A. Gadek [a]

[a] School of Marine Science and Tropical Biology, James Cook University, P.O. Box 6811, Cairns, Qld 4870, Australia
[b] Australian Centre for Evolutionary Biology and Biodiversity, School of Earth and Environmental Science, University of Adelaide, Adelaide, SA 5000, Australia

ARTICLE INFO

*Article history:*
Received 12 August 2008
Revised 12 November 2008
Accepted 13 November 2008
Available online 24 November 2008

*Keywords:*
Secondary structure
ITS
ETS
Sapindaceae
RNA models
Bayesian analyzes

ABSTRACT

Ribosomal DNA internal transcribed spacers (ITS) and partial external transcribed spacers (ETSf) are popularly used to infer evolutionary hypotheses. However, there is generally little consideration given to the secondary structures of these small RNA molecules and their potential effects on sequence alignment and phylogenetic analyzes. Intergeneric relationships amongst three of the four major lineages in the Sapindaceae, the Dodonaeoideae, Hippcastanoideae and Xanthoceroideae were assessed by firstly, generating secondary structure predictions for ITS and partial ETSf sequences, and then these predictions were used to assist alignment of the sequences. Secondly, the alignment was analyzed using RNA specific models of sequence evolution that account for the variation in nucleotide evolution in the independent loops and covariating stems regions of the ribosomal spacers. These models and phylogeny drawn from these analyzes were compared with that from analyzes using 'traditional' 4-state models and previous plastid analyzes. These analyzes identified that paired-site models developed to deal specifically with stem structures in RNA encoding sequences more appropriately account for the evolutionary history of the sequences than traditional 4-state substitution models.



## 1. Introduction

The use of molecular data from the internal transcribed spacers (ITS 1 and ITS 2 = ITS) of the nuclear ribosomal repeat for phylogenetic inference at infrageneric levels, or between closely related genera is widely popular in plant systematics (see Hershkovitz et al., 1999; Alvarez and Wendel, 2003), while the use of a partial fragment 5′external transcribed spacer region (ETSf) at the same taxonomic level is usually dependent on the ease and ability of primer design (Hershkovitz et al., 1999). However, usage of these spacer regions in phylogeny reconstruction is not without concern, particularly with regards to several evolutionary peculiarities that are apparent when sequencing and analyzing these regions. These include non-complete concerted evolution of the ribosomal repeats resulting in intrasequence polymorphisms, preferential sequencing of paralogous non-functional copies (pseudogenes), non-independence of many nucleotide sites due to the properties of secondary structure constraints and the potential for compensatory base changes (CBC), and also noticeable homoplasy, alignment and rooting difficulties due to the faster evolutionary rate of change of nucleotides within spacer sequences (Hershkovitz et al., 1999; Alvarez and Wendel, 2003; Bailey et al., 2003).

In this study potential problems concerning non-independence among sites and associated homoplasy that involves CBC are addressed by using alignment and analyzes methods that incorporate secondary structure (further discussion on these issues below). We also develop a general framework for analyzing small RNA molecules that incorporates testing for functional constraints.

Both ITS and ETSf are part of the cistron that encodes the 18S, 5.8S and 26S single strand rRNAs that occur in tandem arrays located at one or more chromosomal loci. Due to lower selective constraints compared to the coding regions, the primary sequence of these spacers can be highly variable and possibly unalignable between more distantly related species due to length and intrasequence nucleotide heterogeneity between sampled taxa.

rDNAs encode RNA genes, which are single stranded but develop secondary structure where the molecule folds onto itself to form generally short regions of Watson Crick base pairings (G:C and A:U) and the intermediate non-canonical pair (G:U) in stems, and single stranded loops (see Table 1 for glossary). Throughout the rest of this study we use the RNA equivalent U (uracil) for the T (thymine) in the DNA. Stems are generally conserved over evolutionary time and the pairings are maintained by compensatory mutation (CBC–compensatory base change). Knowledge of

* Corresponding author. Fax: +61 7 40421319.
*E-mail address:* Mark.Harrington@jcu.edu.au (M.G. Harrington).

1055-7903/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2008.11.010

the secondary structure of a sequence can provide information in terms of optimal base pairing that can aid alignment (Coleman and Mai, 1997; Denduangboripant and Cronk, 2001). Manual alignment hypotheses that are aided and constrained by secondary structure conventions have been shown to favor phylogenies more congruent with other sources of data than other alignments (see Kjer, 2004 for multiple references).

Conventional phylogenetic analysis methods use models that generally assume that nucleotide sites in a sequence are evolving independently and are spatially distinct. For RNA encoding regions, however, secondary structure dictates that this is not a valid assumption, so that paired-site models developed to deal specifically with stem structures in RNA encoding sequences are more appropriate (Jow et al., 2002; Hudelot et al., 2003). It has also been shown that phylogenetic reconstructions that employ independent assumptions for non-independent data can overestimate support (in terms of bootstrap) for internal branches (Jow et al., 2002; Galtier, 2004; Smith et al., 2004).

Virtually all of the studies utilizing ITS and/or ETSf primary sequences have ignored or barely considered the secondary structure of these molecules. It has been shown that for a range of rRNA sequences the use of models of sequence evolution that allow stem and loop regions to evolve according to separate models significantly improve likelihood-based estimates of phylogeny compared to independent models (Muse, 1995; Schoniger and von Haeseler, 1999; Savill et al., 2001; Telford et al., 2005), and recently this approach has also been shown to be appropriate for use with ITS rDNA sequences (Biffin et al., 2007). There have been no studies that we are aware of that have incorporated secondary structure and paired-sites models for analyzes of ETSf sequences.

Chemical determination of secondary structure is rarely attempted; rather molecular biologists use a comparative biology approach across a sequence alignment to define a putative secondary structure, with the support for the core structure coming from covariation analyzes. A helix is considered 'proven' when it contains at least one CBC, or contains conserved structural motifs or other specific structural elements (e.g. tetraloops—for further examples and references see Gutell et al, 2002). The comparative method assumes that there is a generally maintained secondary structure for a group of sequences, and that the evolutionary processes of selection and mutation do not alter the structure and function of the molecule (Gutell et al., 2002).

This study incorporates an investigation of the secondary structures of the nuclear ribosomal spacers into phylogenetic analyzes of relationships amongst three of the four major lineages in the Sapindaceae, the Xanthoceroideae, Hippocastanoideae and Dodonaeoideae, for which there is near complete sampling. Phylogenetic relationships within Sapindaceae have been previously investigated using a combined plastid dataset of *rbcL* and *matK* genes (Harrington et al., 2005). The plastid gene tree is well resolved in all of the deepest branches; however, there is a lack of resolution within the Dodonaeoideae. Biological knowledge of the evolution of Dodonaeoideae has been difficult to ascertain using morphological and pollen characters, and it has generally been regarded as a heterogeneous assemblage of genera (Radlkofer, 1890; Muller and Leenhouts, 1976; Buijsen et al., 2003). In an attempt to improve the robustness of phylogenetic hypotheses within Dodonaeoideae (Sapindaceae) and to assess congruence with the plastid analyzes, DNA sequences from spacer regions that separate ribosomal genes (rDNA) were generated.

The aims of this study are: (1) to generate secondary structure predictions for ITS and ETSf for Sapindaceae subfamilies Xanthoceroideae, Hippocastanoideae and Dodonaeoideae, (2) to use these predictions to assist alignment of sequences, (3) to use the alignment and the associated structural partitioning mask to select an appropriate model for phylogenetic analysis, and (4) to compare the phylogeny drawn from the analysis with that from plastid data.

## 2. Methods

### 2.1. Sampling, DNA extraction, amplification and sequencing

Twenty-five ITS sequences were generated, with a further six sequences (including two outgroups from Burseraceae) added from GenBank, and 33 ETSf sequences plus two outgroups from GenBank, for a total combined matrix of 40 taxa (Table 2).

Total genomic DNA was isolated from leaf tissue using a CTAB protocol (Doyle and Doyle, 1990), and further cleaned using Jetquick (Genomed). The internal transcribed spacer region of nrDNA (ITS) was amplified using either ABI101/ABI102 (Sun et al., 1994) or ITS4/ITS5 (White et al., 1990) primer pairs, with ETS18S/9 bp (Wright et al., 2001) for ETSf. Amplification and sequencing reactions were as outlined in Harrington and Gadek (2004) except for the addition of 10% dimethyl sulfoxide (DMSO) to some difficult to amplify ITS reactions. Double stranded PCR products were purified with the UltraClean™PCR Clean-up Kit (MO BIO Laboratories Inc., Solana Beach, CA) and sequenced in the forward and reverse direction using dynamic ET dye terminator kit (Megabase) chemistry (Amersham Biosciences). Cleaned PCR products were sequenced on a Megabase 1000 (Amersham Biosciences) at the Genetic Analysis Facility of James Cook University. Forward and reverse sequences were edited with ChromasPro Version 1.32 (Technelysium Pty Ltd.).

### 2.2. Initial alignment and analyzes

All spacer sequences plus the 5.8S gene were initially aligned using the ClustalW function in BioEdit version 7.0.1 (Hall, 1999). The uniform length (164 bp) and very low levels of sequence variation in the 5.8S gene of ITS, along with no substitutions in the ITS 1 motif of Liu and Schardl (1994) or in the three highly conserved 5.8S gene motifs (Jobes and Thien 1997; Harpke and Peterson 2008) and 5.8S gene EcoRV restriction site (Liston et al. 1996), or in key conserved structural motifs in ITS 2 (Liu and Schardl

**Table 1**
Glossary of RNA secondary structure terms (adapted from Gillespie, 2004.)

| Term | Definition |
|---|---|
| Helix (stem) | A double helix composed of a succession of complementary hydrogen-bonded nucleotides between paired strands. Pairing generally involves the Watson–Crick A:U, G:C pairs and the non-canonical G:U pair |
| Single strand loop | Unpaired nucleotides separating helices |
| Terminal loop | Succession of unpaired nucleotides at the end of a stem |
| Lateral bulge | Succession of unpaired nucleotides on one strand of a helix |
| Internal bulge | Group of nucleotides from two parallel strands unable to form canonical pairs |
| Mismatch pair | Any pairing in a secondary structure model that does not involve A:U, G:C or G:U pairs |
| Compensatory base change (CBC) | Subsequent mutation on one strand of a helix to maintain canonical pairing following initial mutation of a complementary base |