Available online at www.sciencedirect.com
ScienceDirect
Molecular Phylogenetics and Evolution 43 (2007) 124–139
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Structural partitioning, paired-sites models and evolution of the ITS transcript in *Syzygium* and Myrtaceae

E. Biffin [a,b,*], M.G. Harrington [c], M.D. Crisp [b], L.A. Craven [a], P.A. Gadek [c]

[a] *Australian National Herbarium, CPBR, CSIRO Plant Industry, GPO Box 1600, Canberra, ACT 2601, Australia*
[b] *School of Botany and Zoology, The Australian National University, Canberra, ACT 0200, Australia*
[c] *School of Tropical Biology, James Cook University, Cairns, Qld 4870, Australia*

Received 17 March 2006; revised 31 July 2006; accepted 14 August 2006
Available online 22 August 2006

Abstract

The internal transcribed spacers (ITS) of nuclear ribosomal DNA are widely used for phylogenetic inference. Several characteristics, including the influence of RNA secondary structure on the mutational dynamics of ITS, may impact on the accuracy of phylogenies estimated from these regions. Here, we develop RNA secondary structure predictions for representatives of the angiosperm family Myrtaceae. On this basis, we assess the utility of structural (stem vs. loop) partitioning, and RNA-specific (paired-sites) models for a 76 taxon *Syzygium* alignment, and for a broader, family-wide Myrtaceae ITS data set. We use a permutation approach to demonstrate that structural partitioning significantly improves the likelihood of the data. Similarly, models that account for the non-independence of stem-pairs in RNA structure have a higher likelihood than those that do not. The best-fit RNA models for ITS are those that exclude simultaneous double substitutions in stem-pairs, which suggests an absence of strong selection against non-canonical (G·U/U·G) base-pairs at a high proportion of stem-paired sites. We apply the RNA-specific models to the phylogeny of *Syzygium* and Myrtaceae and contrast these with hypotheses derived using standard 4-state models. There is little practical difference amongst relationships inferred for *Syzygium* although for Myrtaceae, there are several differences. The RNA-specific approach finds topologies that are less resolved but are more consistent with conventional views of myrtaceous relationships, compared with the 4-state models.

Keywords: ITS1; ITS2; Secondary structure; Stems; Loops; Mutational dynamics; RNA-specific models; Model selection; Maximum likelihood; Bayesian analysis

1. Introduction

Sequences of the internal transcribed spacer (ITS) regions of nuclear ribosomal DNA (rDNA) are a widely used molecular tool for inferring evolutionary relationships amongst eukaryotes (e.g., Hershkovitz and Lewis, 1996; Hershkovitz and Zimmer, 1996; Hershkovitz et al., 1998; Alvarez and Wendel, 2003; Schultz et al., 2005). Several factors, such as high copy number, universality of primer sequences, and the relatively small size of the spacers make data from these regions relatively easy to obtain. In addition, the expectation of high inter-specific and low intra-genomic variability, and bi-parental mode of inheritance has driven the popularity of ITS sequencing (Hershkovitz et al., 1998; Alvarez and Wendel, 2003). While sequencing of ITS has undoubtedly made substantial contributions to phylogenetics, several factors including variable (incomplete) rates of concerted evolution, the presence of divergent pseudogene copies, and highly complex patterns of sequence evolution may confound the reconstruction of historical relationships inferred from these regions (see Alvarez and Wendel, 2003, and references therein).

A specific concern is the influence of RNA secondary structure on the mutational dynamics of ITS, which has important implications for phylogenetic inference (Alvarez and Wendel, 2003). rDNAs encode RNA genes, which are

* Corresponding author.
E-mail address: Ed.Biffin@csiro.au (E. Biffin).

1055-7903/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2006.08.013

single stranded but develop secondary structure (helical regions, or stems, formed by intra-molecular base pairing) as part of the formation and functioning of ribosomes (Noller, 1984). Many RNA molecules are subject to evolutionary constraint related to the maintenance of specific secondary structures that provide functionality. However, it is frequently observed that homologous stable stem structures are maintained despite extensive nucleotide divergence, because stem pairing regions of conserved RNAs evolve via selectively neutral mutations in the form of compensatory (or hemi-compensatory) base pair change (CBC). Mutations at different (often–distant) sites in a molecule can be correlated because a change in one nucleotide in a stem pair must be compensated by a change in the opposing member, in order to preserve an energetically stable secondary structure (Higgs, 2000). Most phylogeny reconstruction methods assume independence amongst sites and may therefore be unsuitable for RNAs with conserved secondary structure (Wheeler and Honeycutt, 1988; Dixon and Hillis, 1993; Tillier and Collins, 1998; Higgs, 2000; Savill et al., 2001; Telford et al., 2005). More generally, there is an expectation that the patterns of evolution may vary substantially between stem-paired and single stranded (loop) regions, and for example, helical regions of RNA molecules tend to be G–C rich suggesting selection to maintain thermodynamically stable stem structures (Hershkovitz et al., 1998; Higgs, 2000; Savill et al., 2001). In single stranded regions, there may be a pronounced bias towards adenine nucleotides, which are associated with several well-characterised RNA structural motifs, some of which are implicated in higher-level (tertiary) structural interactions (Gutell et al., 2000).

The secondary structure of the ITS regions has been estimated for a number of phylogenetic studies, although the focus, primarily, has been the potential of structural information to facilitate homology-based sequence alignment amongst divergent sequences (e.g., Gottschling et al., 2001; Goertzen et al., 2003) or the identification of putative pseudogene copies of the ITS transcript (e.g., Buckler and Holtsford, 1996; Bailey et al., 2003). However, there are no ITS-based studies to date which have attempted to directly incorporate secondary structure information into models of RNA sequence evolution, perhaps in part reflecting the widespread assumption that the ITS are under low functional constraint, and therefore approximate a neutral evolutionary model. Furthermore, and in contrast to rDNA coding regions (such as the 5.8S rDNA gene, see Hershkovitz and Zimmer, 1996), the ITS lack a broad conservation of sequence (e.g., Baldwin, 1992; Hershkovitz and Zimmer, 1996; Hershkovitz et al., 1998) and this is believed to limit the accuracy of conventional approaches to RNA structure prediction (Alvarez and Wendel, 2003). Nevertheless, there is strong evidence for a generally conserved functional role for ITS that is mediated at the sequence and structural level (e.g., Joseph et al., 1999; Côté and Peculis, 2001; Lalev and Nazar, 1999, 2001). The ITS are sequentially cleaved from the large precursor (pre-RNA) molecule (80-90S nucleolar particle) and digested. However, there are close interdependencies in the cleavage pathway, reflecting the need for higher order structure in the pre-RNA, including the ITS, that may be necessary to organise the cleavage sites in close spatial proximity (Lalev et al., 2000). Key structural elements, including cleavage sites and binding sites for nucleolar proteins (including those associated with the spliceosome-like protein complex referred to as the ribosome assembly chaperone, see Lalev et al., 2000), may be essentially conserved across eukaryotes (e.g., van Nues et al., 1994; Mai and Coleman, 1997; Joseph et al., 1999; Lalev and Nazar, 1999; Coleman, 2003; Schultz et al., 2005). Therefore, the concerns relating RNA secondary structural constraints to phylogenetic analysis could reasonably apply to sequences of the ITS regions.

The study of Harrington and Gadek (2004) used ITS sequences to infer evolutionary relationships within the angiosperm genus *Syzygium* and its allies (Myrtaceae) although the hypothesis they present is based upon relatively simple evolutionary models, including maximum parsimony, and a Bayesian analysis employing a model that allows for differential rates of transitions and transversions (HKY85). Phylogenetic studies have demonstrated that structural partitioning and the use of complex evolutionary models may better account for the mutational processes occurring in RNA sequences (Wilgenbusch and De Querioz, 2000; Savill et al., 2001; Jow et al., 2002; Hudelot et al., 2003; Kjer, 2004; Telford et al., 2005). In particular, maximum likelihood (ML) approaches to phylogeny reconstruction have facilitated the development of models of RNA sequence evolution which treat stem nucleotides as paired sites, and thus account for the possible non-independence of sites within stem-pairing regions (e.g., Tillier and Collins, 1998; Schöniger and von Haeseler, 1999; Higgs, 2000; Savill et al., 2001; Jow et al., 2002). Three classes of models (RNA16, RNA7, and RNA6, in the terminology of Savill et al., 2001) provide rates for the commonly observed base-pairs in secondary structure (i.e., Watson–Crick, G·C/C·G/A·U/U·A, and 'wobble', G·U/U·G, pairs) but differ in the treatment of mismatch pairs. RNA16 includes a rate class for each of the possible mismatch pairings (i.e., 16×16 rate matrix), RNA7 includes a single mismatch class (i.e., 7×7 rate matrix; RNA7) while for RNA6, mismatches are completely excluded (i.e., 6×6 rate matrix) from the analysis. Restrictions of these generalised models include those that exclude the possibility of double substitutions (i.e., all double transitions pass through a GU intermediate, and all double transversions pass through a mismatch pair) or enforce base-pair reversal symmetry (e.g., the rate for AU is equal to the rate for UA). Recently, Savill et al. (2001) compared the variants for each class of RNA model for a small-subunit (SSU) rRNA alignment and concluded that the most generalised model from each class best reflects the complexity of RNA evolution.