Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 35 (2005) 569–582
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Taxon sampling effects in molecular clock dating: An example from the African Restionaceae

H. Peter Linder*, Christopher R. Hardy, Frank Rutschmann

Institute for Systematic Botany, University of Zurich, Zollikerstrasse 107, CH-8008 Zurich, Switzerland

Received 12 July 2004; revised 3 December 2004
Available online 21 January 2005

Abstract

Three commonly used molecular dating methods for correction of variable rates (non-parametric rate smoothing, penalized likelihood, and Bayesian rate correction) as well as the assumption of a global molecular clock were tested for sensitivity to taxon sampling. The test dataset of 6854 basepairs for 300 terminals includes a nearly complete sample of the *Restio*-clade of the African Restionaceae (272 of the 288 species), as well as 26 outgroup species. Of this, nested subsets of 35, 51, 80, 120, 150, and the full 300 species were used. Molecular dating experiments with these datasets showed that all methods are sensitive to undersampling, but that this effect is more severe in analyses that use more extreme rate smoothing. Additionally, the undersampling effect is positively related to distance from the calibration node. The combined effect of undersampling and distance from the calibration node resulted in up to threefold differences in the age estimation of nodes from the same dataset with the same calibration point. We suggest that the most suitable methods are penalized likelihood and Bayesian when a global clock assumption has been rejected, as these methods are more successful at finding optimal levels of smoothing to correct for rate heterogeneity, and are less sensitive to undersampling.

Keywords: Molecular dating; NPRS; Penalized likelihood; Bayesian dating; Sampling effects; Restionaceae; Lineage through time plots

1. Introduction

Dating the internal nodes of cladograms is useful for many evolutionary investigations, for example exploring plant–insect co-speciation (e.g. Percy et al., 2004), historical biogeographical analysis (e.g., Conti et al., 2002; Davis et al., 2002; Nagy et al., 2003; Vinnersten and Bremer, 2001), and relating speciation rate changes to palaeo-environmental changes (e.g., Kadereit et al., 2004; Linder, 2003). However, molecular dating is beset by a number of problems. For example, the pseudoprecision and errors that may result from the use of inadequate calibration points, and especially the use of derived calibration points which are not directly based on fossil evidence, have recently received attention (Graur and Martin, 2004; Hedges and Kumar, 2004; Lee, 1999; Shaul and Graur, 2002). Furthermore, the assumption of a global molecular clock has been shown to be invalid in many instances (Gaut, 1998). Various methods have been developed to accommodate rate variation: these include the removal of clades with deviant rates (Takezaki et al., 1995), excluding data-partitions that falsify the clock assumption (Kato et al., 2003), using several local clocks for rate-homogenous clades (i.e., the local clocks approach of Yoder and Yang, 2000), using non-parametric rate smoothing to constrain between internode rate variation (Sanderson, 1997), and searching for the optimal rates using Bayesian methods (Thorne et al., 1998) and penalized likelihood (Sanderson, 2002a). However, there seems to have been no investigation into the effects of sampling only a small proportion of the

* Corresponding author. Fax: +41 1 634 8403.
E-mail address: plinder@systbot.unizh.ch (H.P. Linder).

1055-7903/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2004.12.006

terminals (species) on the age estimates of the interior nodes. An understanding of how undersampling effects age estimates is important, as molecular phylogenetic investigations of clade ages are often based on sparse taxon samples.

Here, we investigate the sensitivity of various methods of obtaining molecular age estimates to incomplete taxon sampling in the "*Restio* clade" of African Restionaceae (Poales) which, with 288 species, is the largest clade of African Restionaceae. The African Restionaceae as a whole comprise 350 species of evergreen, rush-like plants that collectively dominate much of the fynbos vegetation of the species-rich Cape Floristic Region of Southern Africa (Linder, 1991, 2003; Taylor, 1978). Specifically, we evaluate effect on node age estimates of increasing or decreasing taxon sampling, and distance from the calibration node. Our data on the *Restio* clade are particularly suited this type of investigation because (1) taxon sampling is nearly complete (ca. 95%) and (2) phylogenetic relationships are well resolved and supported by over 6000 nucleotides of DNA sequence data.

2. Methods

2.1. Phylogeny estimation

Two hundred and seventy-two species (ca. 95%) of the 288 species of the "*Restio* clade" African Restionaceae were included in the current analysis. Additionally, both subspecies of *Restio dodii* and two accessions of the variable and widespread species *Ischyrolepis macer* were included, as they appear to represent two distinct chloroplast lineages and may be separate species. To allow the use of the basal dating node, we also included 24 species of the "*Willdenowia* clade" of African Restionaceae. As such, a total of 298 plants of African Restionaceae were sampled for this analysis. Of the 16 species of the "*Restio* clade" that were not included, three are possibly not taxonomically distinct (for detailed comment, see Linder, 2001), and the remainder could not be located in the field for the collection of extraction-quality plant material. Based on the phylogenetic studies of Briggs et al. (2000) and Linder et al. (2003), the tree was rooted to two terminals representing the ca. 150 species of Australian Restionaceae.

DNA sequences were generated from the plastid regions spanning the *trnL* intron and the *trnL–trnF* intergenic spacer (Taberlet et al., 1991), the complete gene encoding *rbcL* (Chase and Albert, 1998), the complete *atpB-rbcL* intergenic spacer (Chiang and Schaal, 2000; Cuénoud et al., 2000; Manen et al., 1994), and *matK* plus the flanking *trnK* intron (Hilu and Liang, 1997). Total DNA was isolated from silica gel-dried culms using the DNeasy Plant Mini Kit (Qiagen, Valencia, California, USA). Sequences were generated using standard methods for PCR amplification and automated sequencing.

Raw sequence data files were analysed with the ABI Prism 377 Software Collection 2.1. Contigs were constructed in Sequencher and alignments were performed using the default alignment parameters in Clustal X (Thompson et al., 1997), followed by adjustment by eye. These sequences were assembled into a single matrix in WinClada (Nixon, 2002). The aligned matrix consisted of 6854 aligned bases, of which 1512 are parsimony informative. Additionally, indels were coded at the end of the matrix using Simple Indel Coding (Simmons and Ochoterena, 2000) as implemented in the program GapCoder (Young and Healy, 2001). The total matrix consists of 1782 parsimony–informative characters. All characters were weighted equally and treated as non-additive during tree searches. This data matrix has been deposited at www.treebase.org.

Parsimony searches were conducted using the parsimony ratchet (Nixon, 1999) as implemented from WinClada, running NONA vers. 1.6 (Goloboff, 1993) as a daughter process. Ten ratchet searches were conducted, each initiated with the generation of a Wagner tree, using a random taxon entry sequence, followed by TBR branch swapping with one tree retained and used as the starting point for 500 ratchet cycles. In the weighted/constrained half of each ratchet cycle, a randomly selected set of 10% of the characters were resampled, and a randomly selected set of 10% of the resolved clades were constrained. This analysis resulted in 885 equally most parsimonious cladograms ($L = 5415$, $CI = 0.44$, $RI = 0.84$; informative characters only). These were then pooled and swapped to obtain a total of 10,615 cladograms of length 5415. One of these cladograms was arbitrarily chosen for the subsequent investigation into the impact of taxon sampling on the estimation of absolute dates and divergence times.

2.2. Construction of smaller subset matrices and cladograms

Using our 300 taxon matrix (not including indels) and tree as fixed starting points, six smaller matrices and trees were constructed by deleting terminals in Mesquite 1.02 (Maddison and Maddison, 2003). These smaller datasets have 150, 120, 100, 80, 51, and 35 species/terminals, respectively. The list of species and sequences in each smaller set is a precise subset of the next larger set, and each employed the same relative alignment and tree topology as those obtained from the 300 taxon analysis. The only differences lie in the numbers of terminals and by the exclusion of extraneous gaps from the larger matrices that are no longer necessary in the smaller matrices. As such, each successively smaller matrix consists of 6623, 6547, 6480, 6399, 6248, and 6135 aligned bases. For the smallest (35 species) sampling, at least two