Molecular Phylogenetics and Evolution 49 (2008) 832–842

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow

Andrew J. Eckert [a], Bryan C. Carstens [b,*]

[a] Section of Evolution and Ecology, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA
[b] Department of Biological Sciences, 202 Life Sciences Building, Louisiana State University, Baton Rouge, LA 70803, USA

ARTICLE INFO

Article history:
Received 21 March 2008
Revised 8 September 2008
Accepted 12 September 2008
Available online 21 September 2008

Keywords:
COAL
Gene flow
Estimating species phylogenies
Incomplete lineage sorting
Minimizing deep coalescence
Rhododendron

ABSTRACT

Incomplete lineage sorting has been documented across a diverse set of taxa ranging from song birds to conifers. Such patterns are expected theoretically for species characterized by certain life history characteristics (e.g. long generation times) and those influenced by certain historical demographic events (e.g. recent divergences). A number of methods to estimate the underlying species phylogeny from a set of gene trees have been proposed and shown to be effective when incomplete lineage sorting has occurred. The further effects of gene flow on those methods, however, remain to be investigated. Here, we focus on the performance of three methods of species tree inference, ESP-COAL, minimizing deep coalescence (MDC), and concatenation, when incomplete lineage sorting and gene flow jointly confound the relationship between gene and species trees. Performance was investigated using Monte Carlo coalescent simulations under four models (*n*-island, stepping stone, parapatric, and allopatric) and three magnitudes of gene flow ($N_em = 0.01, 0.10, 1.00$). Although results varied by the model and magnitude of gene flow, methods incorporating aspects of the coalescent process (ESP-COAL and MDC) performed well, with probabilities of identifying the correct species tree topology typically increasing to greater than 0.75 when five more loci are sampled. The only exceptions to that pattern included gene flow at moderate to high magnitudes under the *n*-island and stepping stone models. Concatenation performs poorly relative to the other methods. We extend these results to a discussion of the importance of species and population phylogenies to the fields of molecular systematics and phylogeography using an empirical example from *Rhododendron*.

1. Introduction

The fundamental goal of systematics is to understand the process of lineage divergence that leads to the formation of new species. Since Maddison (1997) there has been growing acceptance among systematists that gene genealogies are not always congruent with species phylogenies (e.g. the actual pattern of lineage splitting and descent from common ancestors). It is now widely recognized that processes such as gene duplication (Fitch, 1970), lateral transfer (Cummings, 1994) and incomplete lineage sorting (Tajima, 1983; Takahata and Nei, 1985; Hudson, 1992) can lead to incongruence between gene trees and species trees, and empirical examples of each process exist (cf. Syring et al., 2007 for an example of incomplete lineage sorting). This realization has prompted the development of approaches designed to estimate species phylogenies despite the process that presumably caused the incongruence. For example, gene tree parsimony (Slowinski and Page, 1999) was developed to account for gene duplication, while the minimization of deep coalescence (MDC; Maddison, 1997), COAL (Degnan and Salter, 2005), and BEST (Edwards et al., 2007; Liu and Pearl, 2007) were designed in part to estimate species phylogeny when the discord between the gene trees and species tree is a result of the incomplete sorting of ancestral polymorphisms.

At the initial stages of divergence, incomplete lineage sorting is ubiquitous and likely produces the majority of gene-species tree discord among closely related lineages. This is a direct outcome of population-level processes; consequently, the developers of methods have incorporated statistical models derived from the coalescent (Kingman, 1982; Hudson, 1990) into species-level phylogenetic analyses to account for these processes. However, for many empirical systems it is also these lineages that exchange migrants, particularly when they occur in sympatry. Since genetic polymorphism shared among lineages can result from either retained ancestral polymorphism or a gene copy introduced into the population via gene flow (Slatkin and Maddison, 1989), it is often difficult to determine which process produced the shared polymorphism. Fully statistical treatments of coalescence, gene flow,

* Corresponding author.
E-mail addresses: carstens@lsu.edu, ajeckert@ucdavis.edu (B.C. Carstens).

1055-7903/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2008.09.008

and divergence are currently available only for pairwise comparisons between two lineages (Nielsen and Wakeley, 2001; Hey and Nielsen, 2004, 2007; Hey, 2006).

It is an understatement to suggest that the biologist who wishes to estimate species phylogeny in a system where details such as (a) the number of lineages, (b) the relationship among lineages, and (c) the amount of gene flow are unclear is currently faced with a difficult task. Methods that estimate a species phylogeny using some approach derived from the coalescent must be robust to at least moderate levels of gene flow (e.g. levels that not be easily recognizable) to be of any use to the majority of empirical biologists, or the use of such methods may result in spurious conclusions about the actual pattern of lineage divergence. The data we present in this manuscript were collected out of a desire to explore how the phylogenetic signal contained in DNA sequence data is affected by gene flow in recently diverged lineages. Does gene flow destroy phylogenetic signal entirely, or are some methods able to accurately estimate species phylogeny when some of the shared polymorphisms result from gene flow? In order to explore this issue, we evaluate approaches based on the coalescent that use estimated gene trees as input in an attempt to isolate gene flow as the sole factor affecting phylogenetic accuracy.

2. Materials and methods

2.1. Statistical inference of species trees from gene trees

A renewed interest exists in the development and interpretation of statistical methods for the inference of species trees from gene trees (Maddison and Knowles, 2006). A myriad of innovative approaches have been developed (Slatkin and Maddison, 1989; Maddison, 1997; Page and Charleston, 1997; Slowinski and Page, 1999; Liu and Pearl, 2006; Edwards et al., 2007; Carstens and Knowles, 2007), as well as applied to empirical questions in phylogeography and systematics (Knowles and Carstens 2007; Brumfield et al., in press; Carling and Brumfield, 2008). Here, we focus on two methods for estimating species phylogenies at relatively low levels of lineage divergence. The first seeks to identify the species tree that maximizes the probability of a set of genealogies given the species tree (Maddison, 1997), as implemented in COAL (Degnan and Salter, 2005) and as applied by Carstens and Knowles (2007). The second method, described by Maddison (1997) and implemented in the Mesquite software package (Maddison and Maddison, 2004) minimizes the amount of deep coalescence to estimate the species phylogeny. Hereafter we refer to these approaches as ESP-COAL and MDC, respectively.

ESP-COAL is a maximum-likelihood approach to the inference of species trees from a set of gene trees. Maddison (1997) noted that the likelihood (L) of a species tree inferred from n independent gene loci could be written as:

$$\mathrm{L(D|ST)} = \prod_{n=1}^{n} \left(\sum_{\mathrm{GT}} [\Pr(D|\mathrm{GT}) \Pr(\mathrm{GT}|\mathrm{ST})] \right) \quad (1)$$

where D are the sequence data, ST is the species tree, and GT is the gene tree. Note that the summation is over all possible GT for each of the n loci. The first expression of the inner product is the likelihood of the data given a gene tree, which can be computed by standard phylogenetic software. The second expression of the inner product represents the probability of a gene tree given a species tree. This quantity can only be calculated for some sample configurations using the mathematical theory for the neutral coalescent (Tajima, 1983, 1989; Hudson, 1983; Takahata, 1989; Rosenberg, 2002; Yang, 2002; Wall, 2003). Degnan and Salter (2005), however, devised a combinatoric approach for the calculation of this probability, with the limitation that it results in the probability of the gene tree topology, not considering branch lengths, conditional on a species tree topology with known branch lengths. Rigorous maximization of the likelihood function would require joint searches through the state space of all possible gene and species tree topologies and their branch lengths using some form of importance sampling (Felsenstein, 2004). In order to approximate the maximization of the likelihood function as defined above, we followed the method of Maddison (1997) and Carstens and Knowles (2007), which searches for the ST topology conferring the highest probability for the observed gene trees.

The second approach to estimating species phylogeny (MDC), also described by Maddison (1997), uses a heuristic search to identify the species phylogeny that minimizes the amount of deep coalescence (e.g. incomplete lineage sorting). This approach can be accurate in the absence of gene flow under certain assumptions concerning the species tree topology (Degnan and Rosenberg, 2006; Maddison and Knowles, 2006), but has not been explicitly explored given varying levels of gene flow. Like the ESP-COAL approach, it evaluates the pattern of coalescence without considering the branch lengths of the genealogies.

2.2. Parameters and models of gene flow

ESP-COAL is accurate when ancestral polymorphisms are segregating within species that otherwise conform to a bifurcating phylogenetic tree (Carstens and Knowles, 2007), particularly when the depth of the species tree is $3N_e$ or greater. However, the signature of ancestral polymorphisms segregating within descendant lineages due strictly to genetic drift is complicated when gene flow, either recent or historical, has occurred among lineages (Slatkin and Maddison, 1989).

We devised four basic models of gene flow in order to elucidate the effects of this process on phylogenetic analyses: n-island, stepping stone, and two models of historical gene flow (Fig. 1). The models of historical gene flow were formulated to reflect scenarios of either allopatric or parapatric speciation. Historical gene flow occurred strictly between sister lineages and was modeled as a burst of gene flow directly after (parapatric) or $0.5xN_e$ generations after speciation (allopatric), where x is the length of time between successive speciation events (Fig. 1). The duration of these bursts was controlled by the parameter d, and we incorporated a relatively short period of divergence with gene flow ($0.1N_e$) as well as a longer period ($0.5N_e$). For each model, we assumed three different magnitudes of gene flow as measured by the effective number of migrants per generation (N_em = 0.01, 0.10, or 1.00).

As shown previously, the power of the ESP-COAL and MDC depend upon the number of unlinked loci used in the analysis and the depth of the species tree (Maddison and Knowles, 2006; Carstens and Knowles, 2007). Therefore, we varied the number of sampled loci from two to ten and the depth of the species tree ($2N_e$ or $6N_e$), as well as the effective population sizes (N_e = 10,000 or 100,000). These parameter treatments were considered in a fully factorial design, yielding 72 different treatment combinations, for each of which we analyzed the accuracy of the ESP-COAL and MDC across samples of two to ten loci (Table S1; online Supplemental data).

2.3. Canonical species tree

We assumed a single, fully resolved, pectinate species tree with four taxa for our simulations [((c:0.75(a:0.375, b:0.375):0.375): 0.25, d:1.00)]. The fourth taxon (taxon d) was designated as an outgroup. The ingroup taxa can then be characterized by three possible rooted tree topologies. In all cases, the relative branch lengths conform to a molecular clock and were defined as pictured in Fig. 1.