Theoretical Population Biology 105 (2015) 17–23

Contents lists available at ScienceDirect

Theoretical Population Biology

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

There are no caterpillars in a wicked forest

James H. Degnan [a,*], John A. Rhodes [b]

[a] Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131, USA

[b] Department of Mathematics and Statistics, University of Alaska Fairbanks, PO Box 756660, Fairbanks, AK 99775, USA

ARTICLE INFO

Article history:
Received 15 May 2015
Available online 10 September 2015

Keywords:
Gene tree
Species tree
Multispecies coalescent
Anomalous gene tree
Coalescent history
Phylogeny

ABSTRACT

Species trees represent the historical divergences of populations or species, while gene trees trace the ancestry of individual gene copies sampled within those populations. In cases involving rapid speciation, gene trees with topologies that differ from that of the species tree can be most probable under the standard multispecies coalescent model, making species tree inference more difficult. Such *anomalous gene trees* are not well understood except for some small cases. In this work, we establish one constraint that applies to trees of any size: gene trees with "caterpillar" topologies cannot be anomalous. The proof of this involves a new combinatorial object, called a *population history*, which keeps track of the number of coalescent events in each ancestral population.

1. Introduction

An important distinction is made in phylogenetics between species trees and gene trees. Species trees describe the ancestral relationships between populations of individuals (each carrying many genes) that have undergone divergences at various times in the past. A gene tree tracks the ancestral relationships for a single gene sampled from individuals within extant species populations. In a species tree, the ancestral populations associated to edges have finite durations (see Fig. 1). As a result, going backwards in time, several gene lineages from sampled individuals may remain distinct within a common ancestral population – a phenomenon called *incomplete lineage sorting* (Maddison, 1997) – and then merge with other lineages to form a gene tree that is topologically dissimilar to the species tree. An understanding of this phenomenon, which leads us to expect some, and possibly many, gene trees to differ from the species tree, is essential to statistical approaches to inference of species trees from genomic data sets.

The multispecies coalescent model gives a stochastic description of gene tree formation within a species tree. Kingman's coalescent model (Kingman, 1982; Hudson, 1983; Tajima, 1983; Wakeley, 2008) is adopted for each population (edge) of the species tree, so that the waiting time until coalescence between any pair of gene lineages within a population, going backwards in time, is exponentially distributed with mean 1. At each node of the species tree, gene lineages reaching it from its descendent edges 'enter' the population above starting a new coalescent process. Combining calculations of probabilities for the within-population Kingman coalescent process with combinatorial features of the species tree, it is possible to calculate the probability of the formation of any topological gene tree (Degnan and Salter, 2005). A rooted species tree, with branch lengths, relating n taxa thus determines a probability mass function on the set of all $(2n-3)!!$ rooted topological gene trees defined on the same species.

Under this model, the most likely gene tree topology does not necessarily match that of the species tree. For example, the species tree $(((a,b),c),d)$, with choices of appropriate branch lengths, can result in any of the symmetric gene tree topologies, $((a,b),(c,d))$, $((a,c),(b,d))$, or $((a,d),(b,c))$, being more probable than the gene tree $(((a,b),c),d)$. The term *anomalous gene tree* (AGT) is used to describe gene trees that are more probable than the gene tree with the same topology as the species tree. Although for four taxa, AGTs only arise for an asymmetric species tree, for any species tree topology with five or more taxa there are branch lengths (durations of internal populations) that lead to at least one AGT (Degnan and Rosenberg, 2006).

Although this result describes the shapes of species trees that can have AGTs, less is known about gene tree shapes that can be AGTs. For four taxa (Degnan and Rosenberg, 2006), explicit computation of gene tree probabilities under the coalescent showed that only symmetric gene trees can be AGTs. For five taxa (Rosenberg and Tao, 2008), a computation showed that if the species tree is

* Corresponding author.
E-mail addresses: jamdeg@unm.edu (J.H. Degnan), j.rhodes@alaska.edu (J.A. Rhodes).

http://dx.doi.org/10.1016/j.tpb.2015.08.007
0040-5809/

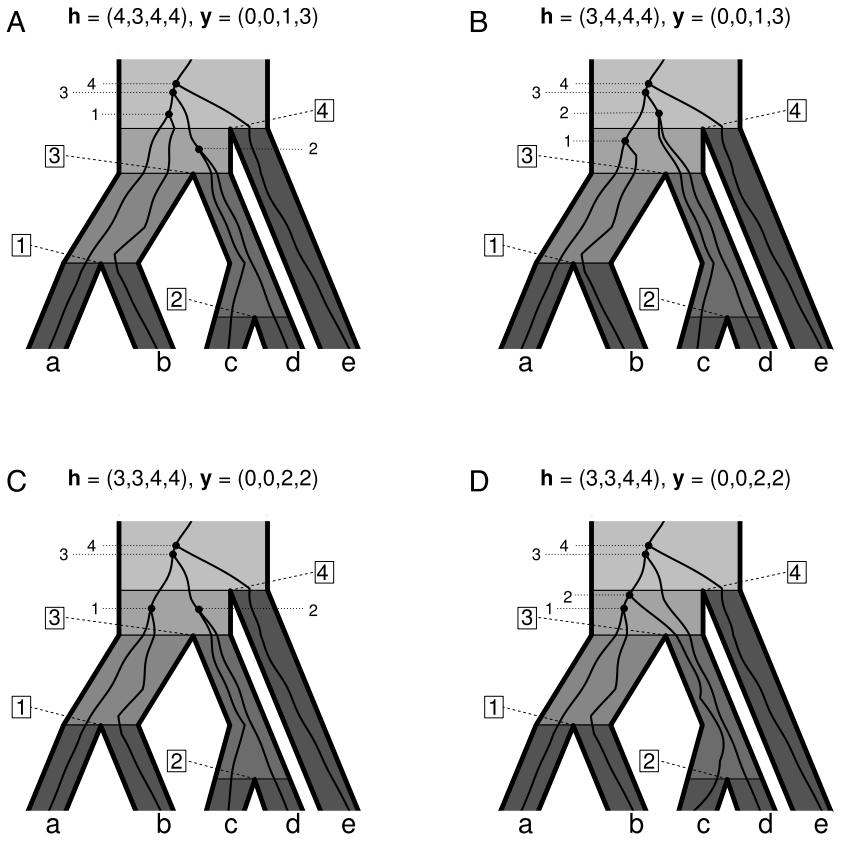

Fig. 1. A species tree with the matching gene tree (A, B, C) under three different coalescent histories (out of 13 possible), and a nonmatching caterpillar gene tree (D). Speciation events occur when populations (shaded polygons) split into two new populations going forward in time (downward). The population ancestral to the root of the species tree (lightest shading) is assumed to extend infinitely into the past; all other populations have finite durations. The nodes of the trees are labeled in a postorder traversal using large, boxed numbers for the species tree, and unboxed numbers for the coalescent events. The vectors **h**, **y** give coalescent histories and population histories, respectively, as explained in Section 2, using node labels as vector indices.

completely unbalanced, e.g., ((((*a*, *b*), *c*), *d*), *e*), then any gene tree with a different unlabeled topology can be an AGT. However, for five-taxon species trees of any topology, a completely unbalanced gene tree is never an AGT. Furthermore, any noncaterpillar gene tree can be an AGT for some species tree. For example, if the species tree is a caterpillar, then any noncaterpillar gene tree is more probable than the matching gene tree if all species tree branch lengths are sufficiently short (Degnan and Rosenberg, 2006).

We refer to completely unbalanced trees, such as ((((*a*, *b*), *c*), *d*), *e*) and its analogs with more taxa, as *rooted caterpillars*, usually omitting the word "rooted" as this paper only concerns rooted trees. We generalize the above observations by showing that for species trees of any size, there are no AGTs with caterpillar topologies. This also implies the statement chosen as the title of this paper, using the terminology introduced in Degnan and Rosenberg (2006) which we restate in the next section.

While our results are theoretical, they have potential to contribute to the practice of species tree inference. For instance, when different genes yield different inferred phylogenetic trees, or different methods yield conflicting estimated species trees, evolutionary biologists sometimes wonder if their inferred tree is an AGT rather than the desired species tree (e.g. Castillo-Ramírez and González, 2008; Zhaxybayeva et al., 2009). A recent paper uses a heuristic test based on taking subsets of four-taxa to conclude that there is evidence of the anomaly zone in a skink phylogeny (Linkem et al., 2014). One implication for our results is that if a phylogenetic method returns a caterpillar tree (as often happens in with smaller numbers of species), the empirical phylogeneticist can be sure that an AGT was not inferred.

2. Notation and definitions

Let X denote a finite set, whose elements we refer to as *taxa*. By a *tree on* X we will mean a topological tree with leaves bijectively labeled by X.

Definition 1. A *species tree* $\sigma = (\psi, \boldsymbol{\lambda})$ *on* X is a rooted, binary tree ψ on X together with a vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{n-2})$ of internal edge lengths (weights), where $n = |X|$, $\{e_1, \ldots, e_{n-2}\}$ are the internal edges of ψ, and $\lambda_i > 0$ is the length of e_i for $i = 1, \ldots, n-2$.

Nodes of the species tree represent speciation events, and edges represent populations extending over time. Edge lengths are given in coalescent units which (for constant population size) are the ratio of elapsed time to population size. It is convenient for the coalescent model to view ψ as augmented by an additional directed edge leading to its root, in order to refer to a population ancestral to the root. We treat this edge as having infinite length, and consider it to be an internal edge of the species tree.

The coalescent on a species tree σ models the formation of gene trees by the merging of ancestral lineages (going backwards in time) within the populations represented by the tree's edges. We focus on the situation where one lineage is sampled per taxon, so pendant edge lengths for the species tree would be irrelevant. With this sampling scheme, a gene tree can also be leaf-labeled by X.

Since under the standard coalescent only binary gene trees have positive probability of being realized, and we are interested solely in the topological form of these trees, we make the following definition.