Inferring incomplete lineage sorting, duplications, transfers and losses with reconciliations

Yao-ban Chan [a,*], Vincent Ranwez [b], Céline Scornavacca [c,d]

[a] *School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia*
[b] *Montpellier SupAgro, UMR AGAP, F-34060 Montpellier, France*
[c] *Institut des Sciences de l'Evolution (ISEM), Université de Montpellier, Montpellier, France*
[d] *Institut de Biologie Computationnelle (IBC), Montpellier, France*

A B S T R A C T

Gene trees and species trees can be discordant due to several processes. Standard models of reconciliations consider macro-evolutionary events at the gene level: duplications, losses and transfers of genes. However, another common source of gene tree-species tree discordance is incomplete lineage sorting (ILS), whereby gene divergences corresponding to speciations occur "out of order". However, ILS is seldom considered in reconciliation models. In this paper, we devise a unified formal $\mathbb{IDTL}$ reconciliation model which includes all the above mentioned processes. We show how to properly cost ILS under this model, and then give a fixed-parameter tractable (FPT) algorithm which calculates the most parsimonious $\mathbb{IDTL}$ reconciliation, with guaranteed time-consistency of transfer events. Provided that the number of branches in contiguous regions of the species tree in which ILS is allowed is bounded by a constant, this algorithm is linear in the number of genes and quadratic in the number of species. This provides a formal foundation to the inference of ILS in a reconciliation framework.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Macro-evolutionary events at the species level (i.e., speciation) impact the genomes of the individuals belonging to the involved species. Hence, the evolutionary history of a group of species strongly influences the evolutionary history of its genes. However, even though species evolution strongly shapes each gene history, it does not fully determine it, and the discrepancy between the two histories provides clues about gene-specific evolutionary events such as gene duplication, gene transfer and gene loss.

Many methods have been proposed to *reconcile* the (inferred) evolutionary history of a gene (depicted as a gene tree) with that of the corresponding species (depicted as a species tree), using gene-specific events. In general, these methods fall into two paradigms: probabilistic methods (e.g., Arvestad et al., 2004; Rasmussen and Kellis, 2011), which find the most likely reconciliation under a statistical model of evolution, and parsimony-based methods (e.g., David and Alm, 2010; Doyon et al., 2011b; Górecki and Tiuryn, 2006), which minimise the number (or total cost under a

penalisation scheme) of the gene-specific events. In this paper, we concentrate on the latter paradigm for reasons of efficiency and scalability.

Gene transfers are particularly difficult to take into account due to the time constraints they induce (Doyon et al., 2011a). Thus, reconciliation methods differ mainly by the way they handle transfer events. Some simply ignore them, relying on the fact that transfers almost never occur in a large part of the animal kingdom (Wu et al., 2014; Zmasek and Eddy, 2001). Some search for optimal reconciliations without considering the time constraints induced by transfers and, if needed, they either modify the inferred solution to satisfy these constraints — with no guarantee of global optimality — or they check for *time-consistency* of the transfers *a posteriori* and return an optimal solution that is time-consistent, but only if any exists (Merkle et al., 2010; Stolzer et al., 2012). Finally, some fully handle transfer events and the associated time constraints in polynomial time, but require that the dates of speciations are provided (David and Alm, 2010; Doyon et al., 2011b; Tofigh et al., 2011).

In addition to discrepancies caused by duplications, transfers and losses, an additional source of discordance between gene and species trees arises from incomplete lineage sorting (Maddison, 1997). In theory, incomplete lineage sorting is not a true "gene
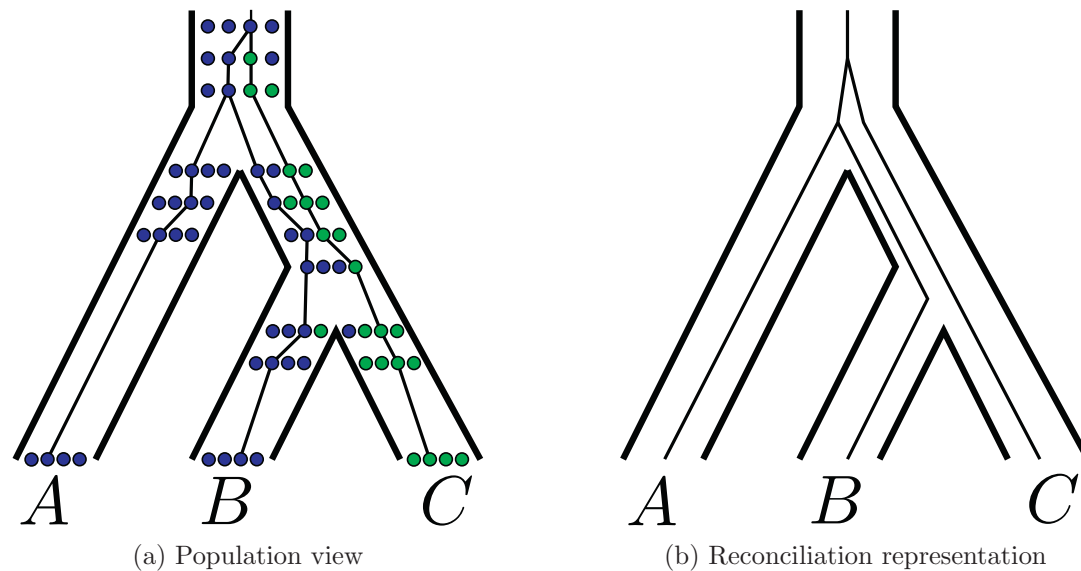
Fig. 1. Impact of incomplete lineage sorting on simple populations of 4 haploid individuals. The originating population contains a single blue allele for the considered gene. First, a mutation leads to a new green allele at this locus, then a first speciation takes place, rapidly followed by a second one. As the blue and green alleles still co-exist when the second speciation takes place, both alleles still have a chance to be fixed in the resulting child species *B* and *C*. For these species, the history of this gene will hence differ from the species history due to ILS. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

event" such as a duplication or a transfer, since nothing "happens" to the gene during incomplete lineage sorting. Still, it is a phenomenon that can lead to a gene tree differing from the species tree containing it. In order to explain how ILS affects gene histories, we recall how a speciation acts on populations.

A speciation can be seen as the division of a population into (two) sub-populations that will evolve separately and hence *fix* potentially different gene variants (alleles) so that those alleles are somehow *sorted* from the originating population in the two sub-populations that eventually become the two new species. For instance, in Fig. 1, the ancestral population giving rise to species *B* and *C*, prior to the speciation, contains blue and green alleles for the considered locus; the speciation leads to two populations, one containing only blue alleles (species *B*) and the other only green alleles (species *C*).

Such a "sorting" is not instantaneous, and if another speciation event occurs soon after the first one, a locus may be *incompletely sorted* at the time of the second speciation. In such a case, we can observe — in the two new species originating from the second speciation — individuals that carry genes whose most recent common ancestor predates the first speciation event. This results in the appearance of the two speciation events being "swapped" in the gene tree, as shown in Fig. 1.

The likelihood of an ILS is mainly related to the ancestral effective population size, which can be hard to estimate, and the time elapsed between the two or more successive speciation events, corresponding to the branch length of a dated species tree. However, in theory, given any species tree, all possible gene tree topologies where each species has exactly one copy of the gene can be explained by ILS alone.

The existence of ILS as a reason for discordance between gene and species trees has been known for some time, and is often used in species tree inference from gene trees (Degnan and Rosenberg, 2009; Liu et al., 2009). In these cases, the multispecies coalescent, arising from Kingman's coalescent in population genetics (Kingman, 1982a; 1982b), provides a statistical model under which the likelihood of ILS can be evaluated.

Inference of ILS via reconciliation is less common. In a seminal paper, Maddison (1997) suggested the parsimonious criterion of minimising deep coalescences (MDC) for reconciliation, where the total number of "extra lineages" in all branches is minimised. An algorithm to solve this problem was constructed by Than and Nakhleh (2009), and extended for the presence of hybridization in Yu et al. (2013).

These papers did not consider macro-events such as duplications and losses, and indeed very few papers attempt to combine both ILS and macro-events in a unified framework. Combining these events is relevant from a biological perspective, as recent studies have shown that ILS and gene introgression (although not specifically LGT) can both occur in the history of a species (Kutschera et al., 2014; McGuire et al., 2007; McLean et al., 2016). More generally, with the increasing availability of data and efficiency of algorithms, the number of species that can be considered in one tree is increasing rapidly. In consequence, even processes which mainly occur in different parts of the species tree must now be considered together, in order to capture all possible processes.

Of interest are a series of papers by Rasmussen and Kellis (2012) and Wu et al. (2014), who model ILS together with duplications and losses using a coalescent model. They devised both a probabilistic algorithm (which was found to be very slow in practice), and a parsimony-based algorithm based on dynamic programming. However, their model does not include transfers and so does not need to consider the associated issues of time-consistency.

Another series of papers by Vernot et al. (2008) and Stolzer et al. (2012) formulated a full model with duplications, transfers, losses and ILS, and devised an algorithm to calculate the most parsimonious reconciliation for this model. Their algorithm starts by contracting short branches of the species tree into multifurcating nodes (polytomies). These are considered the only places where ILS can occur. Since ILS is not penalised in their model, discordance explainable by ILS is always associated to ILS. The remaining discrepancies are then explained by duplications, losses and transfers. However, their treatment of transfers does not guarantee a time-consistent reconciliation; this must be checked *a posteriori* and thus their algorithm may fail to return a solution.

The precise complexity of the problem of finding an optimal reconciliation in the full model with duplication, transfers, losses and ILS is unclear. Firstly, there can be slight variations in the formulation of the model which may have an unknown and poten-