Inferring a Species Tree from Gene Trees under the Deep Coalescence Cost

Louxin Zhang
lxzhang@krdl.org.sg
Kenr Ridge Digital Labs and BioInformatics Centre, National University of Singapore,
21 Heng Mui Kang Terrace, Singapore 119613

Keywords: gene and species trees, gene duplications, coalescency theory, algorithms

1 Introduction

Gene trees are fundamental to molecular systematics. Such trees, constructed from DNA sequences variation at individual genetic loci, are mainly used to infer phylogenies of the taxa that bear these genes. However, because of the presence of paralogy, lineage sorting, and horizontal transfer, gene trees and species trees are often inconsistent (see [1] for example) and so a program that sequences copies of a single gene from various species to reconstruct the species tree can yield an erroneous species tree even if the gene tree is absolutely correct. Hence, a fundamental problem that arises in the study is how to reconcile different gene trees into one species tree [1, 4].

Each event of gene duplications, lineage sorting, and horizontal transfer occurs under different circumstances. For example, deep coalescence events depend on various factors; and they are more likely if the species branches are short (in generations) and wide (in population size) [3]. The lineage sorting was investigated in [3] using coalescence theory. The failure of ancestral gene copies to coalesce (looking backwards in time) into a common ancestral copy until deeper than the previous speciation event is called a *deep coalescence event*. Such a event is illustrated in Figure 1. The *deep coalescence cost* is defined as the number of "extra" gene lineages that fail to coalesce on species branches [3]. For example, the species tree in Figure 1, has a branch with two "extra" gene lineages and a branch with one "extra" gene lineages, for the deep coalescence cost 3. In this work, we mainly investigate the issue of inferring a species tree from gene trees using the deep coalescence cost.

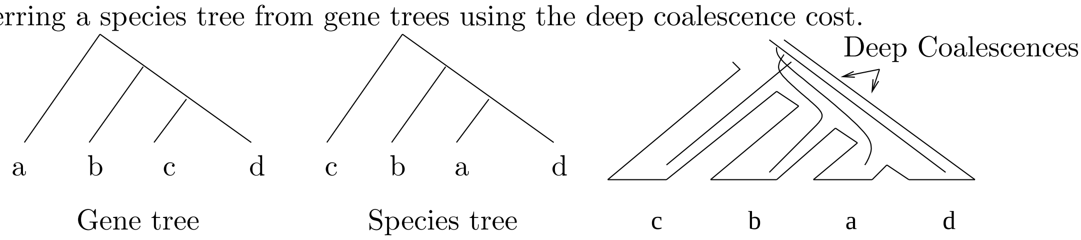

Figure 1: *Deep coalescence events*

2 The deep coalescence cost

For a set I of N biological taxa, the model for their evolutionary history is a rooted full binary tree T where there are N leaves each uniquely labeled by a taxon in I and $N-1$ unlabeled internal nodes. Here the term "full" means that each internal node has exactly two children. Such a tree is called a *species tree*. In a species tree, we treat an internal node as a *cluster* which includes as its members its subordinate species represented by the leaves below it. The model for gene relationship is a rooted full binary tree with labeled leaves. We use the species to label the genes appearing in it. Thus, the labels in a gene tree may not be unique.

Let G be a gene tree and S a species tree such that $L(G) = L(S)$, where $L(G)$ ($L(S)$) denotes the set of leaf labels in the tree. For $s', s'' \in S$, the *least common ancestor* of s' and s'' is defined to be the smallest node $s \in S$ such that $s', s'' \subseteq s$, which is denoted by $\mathrm{lca}(s', s'')$. For any node $g \in G$, we define $M(g)$ as

$$M(g) = \begin{cases} g, & g \in L(G) \\ \mathrm{lca}(M(a(g)), M(b(g))), & g \notin L(G). \end{cases}$$

This correspondence M will be called the *lca mapping* from G to S. Under the lca mapping M from G to S, if a species branch e in S is on k paths from $M(p(g_i))$ to $M(g_i)$, $g_i \in G$ ($1 \leq i \leq k$), then we say that there are $k-1$ "extra" lineages on e and these $k-1$ lineages fails to *coalesce* along e. The *deep coalescence cost* denoted by $t_{dc}(G,S)$ is defined to be the number of "extra" gene branches on species branches [3]. First, we prove a formula that relates the deep coalescence cost to the loss cost, and the duplication cost that were proposed by Page independently [5].

Theorem 2.1 *Let G be a unique leaf-labeled gene tree and S a species tree such that $L(G) = L(S)$. Then, $t_{dc}(G,S) = t_{loss}(G,S) - 2t_{dup}(G,S)$.*

3 Inferring a species tree from gene trees

The problem of inferring a species tree from gene trees can be formulated as finding the most parsimonious species tree by minimizing the deep coalescence cost. Using the technique from [2], we can prove that

Theorem 3.1 *Inferring a species tree from gene trees is NP-hard under the deep coalescence cost.*

Since inferring a species tree from gene trees is NP-hard, there is unlikely a polynomial time algorithm for it unless $NP = P$. Since the duplication cost is much smaller than the loss cost when the gene trees are large, the deep coalescence cost is approximately equal to the loss cost by Theorem 2.1. Therefore, we have implemented the following heuristic algorithm for the problem based on the study on the loss cost ([2]):

Input: A set of gene tree $G_1, G_2, \cdots, G_n$. 1. Use the heuristic algorithm in [2] to find optimal species trees T under the loss lost. 2. Search for optimal species trees from T using alternate nearest neighbor interchange and CP (cut and paste).

References

[1] Fitch, W.M., Distinguishing homologous from analogous proteins, *Syst. Zool.*, 19(2):99–113, 1970.

[2] Ma, B., Li, M., and Zhang, L., From gene trees to species trees, To appear on *SIAM Journal on Computing.* Its extended abstract appeared on *Recomb'98*, 182–191, New York, 1998.

[3] Maddison, W. P., Gene trees in species trees, *Syst. Biol.* 46:523–536, 1997.

[4] Nei, M., Molecular Evolutionary Genetics, Columbia University Press, New York, 1987.

[5] Page, R., Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas, *Syst. Biol.* 43:58–77, 1994.

[6] Myers, G., Selznick, S., Zhang, Z. and Miller, W., Progressive multiple alignment with constraints, *J. Comput. Biol.*, 3(4):563–572, 1996.