CHAPTER 1

Gene Trees, Species Trees, and Species Networks

Luay Nakhleh Derek Ruths
Department of Computer Science
Rice University
Houston, TX 77005, USA
{nakhleh,druths}@cs.rice.edu

Hideki Innan
University of Texas Health Science Center at Houston
Houston, TX 77030, USA
Hideki.Innan@uth.tmc.edu

1.1 Introduction

The availability of whole-genome sequence data has provided a rich resource of deep insights into many biological, medical and pharmaceutical problems and applications, and is promising even more. Yet, along with these insights and promises, genomic data have given rise to many challenging problems, mainly due to the quantity and heterogeneity of such data. One of these major challenges is the phylogenetic analysis of multiple gene datasets that whole genomes provide.

Phylogeny, i.e., the evolutionary history of a set of organisms, has become an indispensable tool in the post-genomic era. Emerging techniques for handling essential biological tasks (e.g., gene finding, comparative genomics, and haplotype inference) are usually guided by an underlying phylogeny. The performance of these techniques, therefore, depends heavily on the quality of the phylogeny. Almost all phylogenetic methods, however, assume that evolution is a process of strict divergence that can be modeled by a phylogenetic tree. While the tree model gives a satisfactory first-order approximation for many families of organisms, other families exhibit evolutionary events that cannot be represented by a tree. In particular, the evolutionary history of bacterial genomes is characterized by the occurrence of processes such as horizontal gene transfer (HGT)—transfer of genetic material across the boundaries of of distantly related species—and inter-specific recombination—exchange of genetic material. Further, hybrid speciation occurs among various groups of plants, fish, and

frogs. In the presence of such evolutionary processes, the evolutionary relationship of a set of organisms is modeled by a *phylogenetic network*.

Accurate reconstruction of these processes bears significant impact on many domains. The Tree of Life—the phylogeny of all organisms on Earth—is one of the grand challenges in evolutionary biology. The prokaryotic branch of this tree is believed to have a large number of horizontal gene transfer events, in addition to recombination events. Efforts to reconstruct a phylogeny for the prokaryotic branch may prove futile without developing phylogenetic network models and reconstruction methods.

A significant aspect of these complex evolutionary mechanisms is their contribution to microbial genome diversification. Like all forms of life, bacteria undergoes evolution. However, unlike many other organisms, bacterial evolution is not one of strict divergence. Recombination usually occurs within populations; in bacteria, however, recombination occurs among different strains. Further, HGT is ubiquitous in the prokaryotic branch of the Tree of Life. Ho (2002) has recently written of the various health risks that recombination and HGT pose, including: (1) antibiotic resistance genes spreading to pathogenic bacteria, (2) disease-associated genes spreading and recombining to create new viruses and bacteria that cause diseases, and (3) transgenic DNA inserting into human cell, triggering cancer. Hence, detecting and reconstructing these processes in bacteria play a major role in developing effective antibiotics, and bears a great impact on human health.

Biologists have long acknowledged the presence of these processes, their significance, and their effects. The computational research community has responded in recent years and proposed a plethora of methods for reconstructing complex evolutionary histories. The general theme of most existing methods can be summarized by: construct gene trees and reconcile them (this is known as the *separate analysis* approach). Gene tree reconciliation presents two major issues, namely identifying the (biological) source of incongruence, and (computationally) reconciling the trees. Many processes may lead to *incongruent* gene trees:
(1) *Stochastic factors*, such as wrong assumptions, insufficient data, incomplete sampling, and differential rates of sequence evolution across lineages. These factors do not violate the tree model of organismal evolutionary relationships; rather, the incongruence they cause must be eliminated in the early stages of phylogenetic analyses.
(2) *Intra-species factors,* such as gene loss and duplication. Although these events may lead to incongruent gene trees, they do not violate the tree model of organismal evolutionary relationships.
(3) *Inter-species factors,* such as horizontal gene transfer (whose rate is very high among prokaryotic organisms), and inter-specific recombination. These events result in *networks* of relationships, rather than trees of relationships.

In this work, we review the intra- and inter-species factors that cause gene tree incongruence and discuss current approaches for resolving these phenomena, with focus on non-treelike evolution. Further, we address extensions to the *coalescent* model to address non-treelike evolution. The rest of the chapter is organized as follows. In

Section 1.2 we illustrate some of the processes that lead to incongruence gene trees. In Section 1.3 we review existing methods for addressing gene tree incongruence caused by gene loss and duplication (intra-species factors). In Section 1.4, we describe the *phylogenetic network* model and discuss the problem of reconciling gene trees into species networks. In Section 1.5 we propose approaches for extending the coalescent model to incorporate non-treelike evolutionary processes. We conclude the chapter in Section 1.6.

1.2 Gene Tree Incongruence

A **gene tree** is a model of how a gene evolves through duplication, loss, and nucleotide substitution. As a gene at a locus in the genome replicates and its copies are passed on to more than one offspring, branching points are generated in the gene tree. Because the gene has a single ancestral copy, barring recombination, the resulting history is a branching tree (Maddison (1997)). Sexual reproduction and meiotic recombination within populations break up the genomic history into many small pieces, each of which has a strictly treelike pattern of descent (Hudson (1983b); Hein (1990); Maddison (1995)). Thus, within a species, many tangled gene trees can be found, one for each nonrecombined locus in the genome. A **species tree** depicts the pattern of branching of species lineages via the process of speciation. When reproductive communities are split by speciation, the gene copies within these communities likewise are split into separate bundles of descent. Within each bundle, the gene trees continue branching and descending through time. Thus, the gene trees are contained within the branches of the species phylogeny (Maddison (1997)).

Gene trees can differ from one another as well as from the species tree. Disagreements (incongruence) among gene trees may be an artifact of the data and/or methods used (stochastic factors). Various studies show the effects of stochastic factors on the performance of phylogenetic tree reconstruction methods (e.g., Hillis *et al.* (1993); Hillis & Huelsenbeck (1994, 1995); Nakhleh *et al.* (2001a,b, 2002); Moret *et al.* (2002)). Stochastic factors confound the accurate reconstruction of evolutionary relationships, and must be handled in the first stage of a phylogenetic analysis. Incongruence among gene trees due to intra- or inter-species processes, on the other hand, is a reflection of true biological processes, and must be handled as such.

Whereas eukaryotes evolve mainly though lineal descent and mutations, bacteria obtain a large proportion of their genetic diversity through the acquisition of sequences from distantly related organisms, via horizontal gene transfer (HGT) or recombination (Ochman *et al.* (2000)). Views as to the extent of HGT and recombination in bacteria vary between the two extremes, with most views being in the middle (Doolittle (1999b,a); Kurland *et al.* (2003); *et al.* (2002); Hao & Golding (2004); *et al.* (2004); Nakamura *et al.* (2004)). However, there is a common belief that recombination and HGT, among other processes, form the essence of prokaryotic evolution. Further, these two are the main processes (in addition to random mutations) by which bacteria develop resistance to antibiotics (e.g., Lewis (1995); Ho (2002);

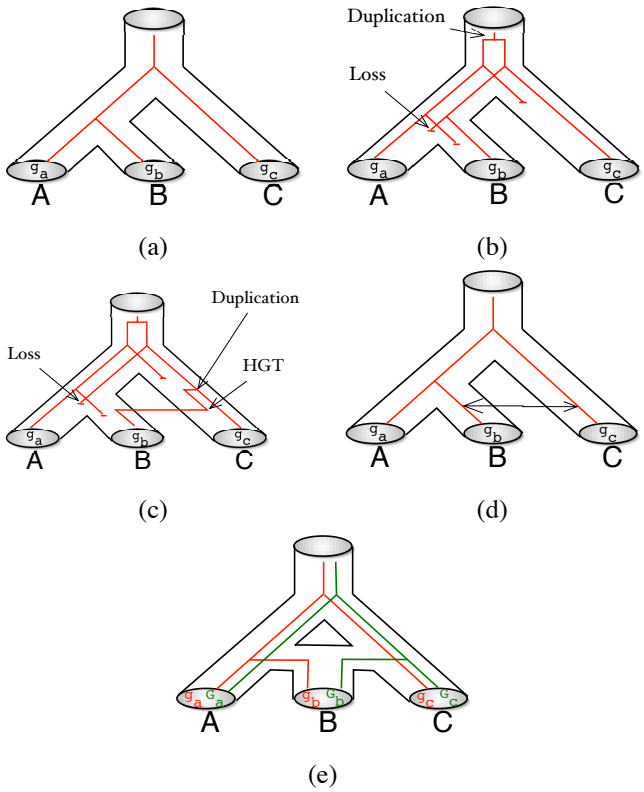


Figure 1.1 *(a) Gene tree that agrees with the species tree. (b) Gene tree that disagrees with the species tree due to gene loss and duplication. (c) Gene tree that disagrees with the species tree due to HGT. (d) An inter-specific recombination event in which genetic material is exchanged between species B and C. (e) A hybrid speciation event that leads to two incongruent gene trees.*

Enright *et al.* (2002); Paulsen *et al.* (2003)). Gene transfer and exchange are considered a primary explanation of incongruence among bacterial gene phylogenies and a significant obstacle to reconstructing the prokaryotic branch of the Tree of Life (Daubin *et al.* (2003)).

We illustrate some of the scenarios that may lead to gene tree incongruence in Figure 1.1. The species (or, organismal) tree is represented by the "tubes"; it has A and B as sister taxa whose most recent common ancestor (MRCA) is a sister taxon of C. Figure 1.1(a) shows a gene evolving within the branches of the same species tree; in this case, the topologies of the gene and species trees agree (the topology of this gene tree is shown in Figure 1.2(a)). In Figure 1.1(b) we show an example of how intra-species processes may lead to incongruent gene trees. The figure shows a gene evolving within the branches a species tree with one duplication event and three losses. Note that the species tree differs from the gene tree; based on this gene, B

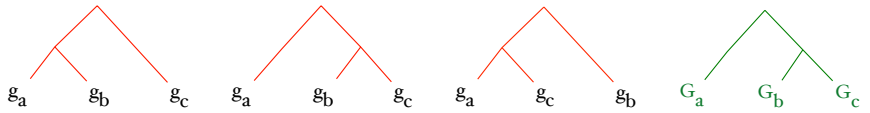


Figure 1.2 *(a) The tree of the gene whose evolution is shown in Figure 1.1(a), and Figure 1.1(e). (b) The tree of the genes whose evolution is shown in Figures 1.1(b) and 1.1(c). (c) The tree of the gene involved in the recombination event shown in Figure 1.1(d). (e) The tree of the gene involved in the hybrid speciation event shown in Figure 1.1(e).*

and C are sister taxa and their MRCA is a sister of taxon A. This gene tree is shown in Figure 1.2(b).

Another event that may cause incongruence between the species tree and the gene tree is HGT. In the case of HGT, shown in Figure 1.1(c), genetic material is transferred from one lineage to another. Sites that are not involved in a horizontal transfer are inherited from the parent (as in Figure 1.2(a)), while other sites are horizontally transferred from another species (as in Figure 1.2(b)).

In the case of inter-specific recombination, as illustrated in Figure 1.1(d), some genetic material is exchanged between pairs of species; in this example, species B and C exchange genetic material. The genes involved in this exchange have an evolutionary history (shown in Figure 1.2(c)) that is incongruent with that of the species. In hybrid speciation, two lineages recombine to create a new species. We can distinguish *diploid hybridization*, in which the new species inherits one of the two homologs for each chromosome from each of its two parents—so that the new species has the same number of chromosomes as its parents, and *polyploid hybridization*, in which the new species inherits the two homologs of each chromosome from both parents—so that the new species has the sum of the numbers of chromosomes of its parents. Under this last heading, we can further distinguish *allopolyploidization*, in which two lineages hybridize to create a new species whose ploidy level is the sum of the ploidy levels of its two parents (the expected result), and *auto-polyploidization*, a regular speciation event that does not involve hybridization, but which doubles the ploidy level of the newly created lineage. Prior to hybridization, each site on each homolog has evolved in a tree-like fashion, although, due to meiotic recombination, different strings of sites may have different histories. Thus, each site in the homologs of the parents of the hybrid evolved in a tree-like fashion on one of the trees induced by (contained inside) the network representing the hybridization event. Figure 1.1(e) shows a network with one hybrid. Each site evolves down exactly one of the two trees shown in Figures 1.2(a) and 1.2(d).

Notice that in the case of intra-species processes (gene loss and duplication), inferring the species tree from a set of potentially conflicting gene trees is a problem of *reconciling* the gene trees and explaining their differences through duplications and losses of genes. Therefore, in this case, despite the potential incongruence among the species and gene trees, the species phylogeny is still a tree (Mirkin *et al.* (1995); Page

(1990, 1994); Eulenstein *et al.* (1998)). However, in the case of recombination, HGT, and hybrid speciation, the evolutionary history of the organismal genomes cannot be represented by phylogenetic trees; rather, *phylogenetic networks* are the appropriate model (Hallett & Lagergren (2001); Moret *et al.* (2004)).

1.3 Trees Within Trees: The Gene Tree Species Tree Problem

Various reports of instances and effects of gene loss and duplication exist in the literature (e.g., Moore (1995); Nichols (2001); Ruvolo (1997)). When losses and duplications are the only processes acting on the genes, a mathematical formulation of the gene tree reconciliation problem is as follows:

Definition 1.1 *(The Gene Tree Reconciliation Problem)*

> ***Input:*** *Set $\mathcal{T}$ of rooted gene trees, a cost w_D for duplications, and a cost w_L for losses.*
>
> ***Output:*** *Rooted tree T with each gene tree $t \in \mathcal{T}$ mapped onto T, so as to minimize the sum $w_D n_D + w_L n_L$, where n_D is the total number of duplications and n_L is the total number of losses, over all genes.*

This problem was shown to be NP-hard by Fellows *et al.* (1998) and Ma *et al.* (1998). Heuristics for the problem exist, but these do not solve the optimization problem (see Ma *et al.* (1998); Page & Charleston (1997a)). Various fixed-parameter approaches have been proposed by Stege (1999a); Hallett & Lagergren (2000) and some variants can be approximated to within a factor of 2 and shown by Ma *et al.* (1998).

When loss and duplication are the only processes acting on the genes, two different questions can be posed, depending on the input data:

1. Gene tree reconciliation problem—when the gene trees are known and the species tree is known, what is the best set of duplication and loss events that reconcile each gene tree with the species tree?
2. Species tree construction problem—when the gene trees are known, but the evolutionary relationships among the species involved is not known, can the gene trees provide the information necessary to derive an estimate of the species tree?

Both of these questions require the assumption of a certain model of gene duplication and loss. The complexity of the gene-tree reconciliation problem is determined by the model chosen, whereas the general species tree construction problem is NP-hard under all commonly used models of gene duplication and loss.

The simplest version of either problem uses a duplication-only model (i.e., losses do not occur). During the period between years 1995 and 2000, this was a commonly used model (Eulenstein *et al.* (1996); Page & Charleston (1997b); Page (1998); Eulenstein (1997); Stege (1999b); Ma *et al.* (1998); Zhang (1997); Ma *et al.* (2000)).

Under the duplication-only model, the gene tree reconciliation problem has linear-time solutions (Zhang (1997); Eulenstein (1997)), as well as other polynomial-time algorithms that report better performance on real biological datasets (Zmasek & Eddy (2001)). The species tree construction problem is NP-hard , as was shown by Ma *et al*. (1998). Different approaches have been taken to solving the species tree construction problem including heuristics (Page & Charleston (1997b)), approximation algorithms (Ma *et al*. (2000)), and fixed parameter tractable algorithms obtained by parameterizing by the number of gene duplications separating a gene tree from the species tree (Stege (1999b)).

The other common model used is the more general duplication-loss model, which admits both duplication and loss events within gene trees. The gene tree reconciliation problem has been shown to be polynomial-time under conditions where the evolution of the sequences themselves are not considered (Arvestad *et al*. (2004); Chen *et al*. (2000); Durand *et al*. (2005)); if this is taken into account, the problem becomes NP-hard (Fellows *et al*. (1998); Ma *et al*. (1998)). Various efficient heuristics for the problem are currently available (Arvestad *et al*. (2003, 2004)). Early work on the gene tree reconciliation problem under this model borrowed techniques from biogeography and host/parasite evolution (Charleston (2000); Page & Charleston (1998)).

1.4 Trees Within Networks: The Gene Tree Species Network Problem

As described in Section 1.2, when events such as horizontal gene transfer, hybrid speciation, or recombination occur, the evolutionary history can no longer be modeled by a tree; rather, *phylogenetic networks* are the appropriate model in this case. In this section, we describe the phylogenetic network model and approaches for reconstructing networks from gene trees.

1.4.1 Terminology and notation

Given a (directed) graph G, let $E(G)$ denote the set of (directed) edges of G and $V(G)$ denote the set of nodes of G. Let (u, v) denote a directed edge from node u to node v; u is the *tail* and v the *head* of the edge and u is a *parent* of v. The *indegree* of a node v is the number of edges whose head is v, while the *outdegree* of v is the number of edges whose tail is v. A node of indegree 0 is a *leaf* (often called a *tip* by systematists). A directed path of length k from u to v in G is a sequence $u_0u_1 \cdots u_k$ of nodes with $u = u_0$, $v = u_k$, and $\forall i$, $1 \leq i \leq k$, $(u_{i-1}, u_i) \in E(G)$; we say that u is the tail of p and v is the head of p. Node v is *reachable* from u in G, denoted $u \rightsquigarrow v$, if there is a directed path in G from u to v; we then also say that u is an *ancestor* of v. A *cycle* in a graph is a directed path from a vertex back to itself; trees never contain cycles: in a tree, there is always a unique path between two distinct vertices. Directed acyclic graphs (or DAGs) play an important role on our model; note that every DAG contains at least one vertex of indegree 0. A *rooted*

directed acyclic graph, in the context of this paper, is then a DAG with a single node of indegree 0, the *root*; note that all all other nodes are reachable from the root by a (directed) path of graph edges. We denote by $r(T)$ the root of tree T and by $L(T)$ the leaf set of T. Let T be a rooted phylogenetic tree over set S of taxa, and let $S' \subseteq S$. We denote by $T(S')$ the minimal rooted subtree of T that connects all the element of S'. Furthermore, the restriction of T to S', denote $T|S'$ is the rooted subtree that is obtained from $T(S')$ by suppressing all vertices (except for the root) whose number of incident edges is 2. Let S' be a maximum-cardinality set of leaves such that $T_1|S' = T_2|S'$, for two trees T_1 and T_2; we call $T_1|S'$ (equivalently, $T_2|S'$) the maximum agreement subtree of the two trees, denoted $MAST(T_1, T_2)$. A *clade* of a tree T is a complete subtree of T. Let $T' = MAST(T_1, T_2)$; then, $T_1 - T'$ is the set of all maximal clades whose pruning from T_1 yields T' (we define $T_2 - T'$ similarly). In other words, there do not exist two clades u and u' in $T_1 - T'$ such that either u is a clade in u', or u' is a clade in u.

We say that node x reaches node y in tree T if there is a directed path from x to y in T. We denote the root of a clade t by $r(t)$. We say that clade t_1 reaches clade t_2 (both in tree T) if $r(t_1)$ reaches $r(t_2)$. The sibling of node x in tree T is node y, denoted $sibling_T(x) = y$ whenever x and y are children of the same node in T. We denote by T_x the clade rooted at node x in T. The least common ancestor of a set X of taxa in tree T, denoted $lca_T(X)$ is the root of the minimal subtree of T that contains the leaves of X. The edge incoming into node x in tree T is denoted by $inedge_T(x)$.

1.4.2 Phylogenetic networks

Moret *et al.* (2004) modeled phylogenetic networks using directed acyclic graphs (DAGs), and differentiated between "model" networks and "reconstructible" ones.

Model networks A phylogenetic network $N = (V, E)$ is a rooted DAG obeying certain constraints. We begin with a few definitions.

Definition 1.2 *A node $v \in V$ is a* tree node *if and only if one of these three conditions holds:*

- $indegree(v) = 0$ *and* $outdegree(v) = 2$*:* root*;*
- $indegree(v) = 1$ *and* $outdegree(v) = 0$*:* leaf*; or*
- $indegree(v) = 1$ *and* $outdegree(v) = 2$*:* internal tree node.

A node v is a network node *if and only if we have* $indegree(v) = 2$ *and* $outdegree(v) = 1$.

Tree nodes correspond to regular speciation or extinction events, whereas network nodes correspond to reticulation events (such as hybrid speciation and horizontal gene transfer). We clearly have $V_T \cap V_N = \emptyset$ and can easily verify that we have $V_T \cup V_N = V$.

Definition 1.3 *An edge $e = (u, v) \in E$ is a* tree edge *if and only if v is a tree node; it is a* network edge *if and only if v is a network node.*

The tree edges are directed from the root of the network towards the leaves and the network edges are directed from their tree-node endpoint towards their network-node endpoint.

A phylogenetic network $N = (V, E)$ defines a partial order on the set V of nodes. We can also assign times to the nodes of N, associating time $t(u)$ with node u; such an assignment, however, must be consistent with the partial order. Call a directed path p from node u to node v that contains at least one tree edge a *positive-time directed path*. If there exists a positive-time directed path from u to v, then we must have $t(u) < t(v)$. Moreover, if $e = (u, v)$ is a network edge, then we must have $t(u) = t(v)$, because a reticulation event is effectively instantaneous at the scale of evolution; thus reticulation events act as synchronization points between lineages.

Definition 1.4 *Given a network N, two nodes u and v cannot* co-exist *(in time) if there exists a sequence $P = \langle p_1, p_2, \ldots, p_k \rangle$ of paths such that:*

- *p_i is a positive-time directed path, for every $1 \leq i \leq k$;*
- *u is the tail of p_1, and v is the head of p_k; and*
- *for every $1 \leq i \leq k - 1$, there exists a network node whose two parents are the head of p_i and the tail of p_{i+1}.*

Obviously, if two nodes x and y cannot co-exist in time, then a reticulation event between them cannot occur.

Definition 1.5 *A* model phylogenetic network *is a rooted DAG obeying the following constraints:*

1. *Every node has indegree and outdegree defined by one of the four combinations $(0, 2)$, $(1, 0)$, $(1, 2)$, or $(2, 1)$—corresponding to, respectively, root, leaves, internal tree nodes, and network nodes.*
2. *If two nodes u and v cannot co-exist in time, then there does not exist a network node w with edges (u, w) and (v, w).*
3. *Given any edge of the network, at least one of its endpoints must be a tree node.*

Reconstructible networks Definition 1.5 of model phylogenetic networks assumes that complete information about every step in the evolutionary history is available. Such is the case in simulations and in artificial phylogenies evolved in a laboratory setting—hence our use of the term *model*. When attempting to reconstruct a phylogenetic network from sample data, however, a researcher will normally have only incomplete information, due to a combination of extinctions, incomplete sampling, and abnormal model conditions. Extinctions and incomplete sampling have the same consequences: the data do not reflect all of the various lineages that contributed to the current situation. Abnormal conditions include insufficient differentiation along

edges, in which case some of the edges may not be reconstructible, leading to polytomies and thus to nodes of outdegree larger than 2. All three types of problems may lead to the reconstruction of networks that violate the constraints of Definition 1.5. (The distinction between a model phylogeny and a reconstructible phylogeny is common with trees as well: for instance, model trees are always rooted, whereas reconstructed trees are usually unrooted. In networks, both the model network and the reconstructed network must be rooted: reticulations only make sense with directed edges.) Clearly, then, a reconstructible network will require changes from the definition of a model network. In particular, the degree constraints must be relaxed to allow arbitrary outdegrees for both network nodes and internal tree nodes. In addition, the time coexistence property must be reconsidered.

There are at least two types of problems in reconstructing phylogenetic networks. First, slow evolution may give rise to edges so short that they cannot be reconstructed, leading to polytomies. This problem cannot be resolved within the DAG framework, so we must relax the constraints on the outdegree of tree nodes. Secondly, missing data may lead methods to reconstruct networks that violate indegree constraints or time coexistence. In such cases, we can postprocess the reconstructed network to restore compliance with most of the constraints in the following three steps:

1. For each network node w with outdegree larger than 1, say with edges (w, v_1), $\ldots$, (w, v_k), add a new tree node u with edge (w, u) and, for each i, $1 \leq i \leq k$, replace edge (w, v_i) by edge (u, v_i).
2. For each network node w whose parents u and v violate time coexistence, add two tree nodes w_u and w_v and replace the two network edges (u, w) and (v, w) by four edges: the two tree edges (u, w_u) and (v, w_v) and the two network edges (w_u, w) and (w_v, w).
3. For each edge (u, v) where both u and v are network nodes, add a new tree node w and replace the edge (u, v) by the two edges (u, w) and (w, v).

The resulting network is consistent with the original reconstruction, but now satisfies the outdegree requirement for network nodes, obeys time coexistence (the introduction of tree edges on the paths to the network node allow arbitrary time delays), and no longer violates the requirement that at least one endpoint of each edge be a tree node. Moreover, this postprocessing is unique and quite simple. We can thus define a reconstructible network in terms similar to a model network.

Definition 1.6 *A* reconstructible phylogenetic network *is a rooted DAG obeying the following constraints:*

1. *Every node has indegree and outdegree defined by one of the three (indegree,outdegree) combinations* $(0, x)$, $(1, y)$, *or* $(z, 1)$, *for* $x \geq 1$, $y \geq 0$, *and* $z \geq 2$*—corresponding to, respectively, root, other tree nodes (internal nodes and leaves), and network nodes.*
2. *If two nodes* u *and* v *cannot co-exist in time, then there does not exist a network node* w *with edges* (u, w) *and* (v, w).
3. *Given any edge of the network, at least one of its endpoints must be a tree node.*

Definition 1.7 *A network N induces a tree T' if T' can be obtained from N by the following two steps:*

1. *For each network node in N, remove all but one of the edges incident into it; and*
2. *for every node v such that $indegree(v) = outdegree(v) = 1$, the parent of v is u, and the child of v is w, remove v and the two edges (u, v) and (v, w), and add new edge (u, w) (this is referred to in the literature as the* forced contraction *operation).*

For example, the network N shown in Figure 1.1(e) induces both trees shown in Figure 1.2(a) and Figure 1.2(d).

1.4.3 Reconstructing networks from gene trees

From a graph-theoretic point of view, the problem can be formulated as pure phylogenetic network reconstruction (Moret *et al.* (2004); Nakhleh *et al.* (2004, 2005)). In the case of HGT, and despite the fact the evolutionary history of the set of organisms is a network, Lerat *et al.* (2003) showed that an underlying species tree can still be inferred. In this case, a phylogenetic network is a pair (T, Ξ), where T is the species (organismal) tree, and Ξ is a set of HGT edges whose addition to T results in a phylogenetic network N that induces all the gene trees. The problem can be formulated as follows.

Definition 1.8 *(The HGT Reconstruction Problem)*

Input: *A species tree ST and a set G of gene trees.*

Output: *Set Ξ of minimum cardinality whose addition to ST yields phylogenetic network N that induces each of the gene trees in G.*

However, in the case of hybrid speciation, there is no underlying species tree; instead the problem is one of reconstructing a phylogenetic network N that induces a given set of gene trees.

Definition 1.9 *(The Hybrid Speciation Reconstruction Problem)*

Input: *A set G of gene trees.*

Output: *A Phylogenetic network N with minimum number of network nodes that induces each of the gene trees in G.*

The minimization criterion reflects the fact that the simplest solution is sought; in this case, the simplest solution is one with the minimum number of HGT or hybrid speciation events. We illustrate this point with the example species tree ST in Figure 1.3(a) and the gene tree GT in Figure 1.3(b). Assume that the actual HGT events that took place are the one depicted in Figure 1.3(c). Nonetheless, the scenario depicted

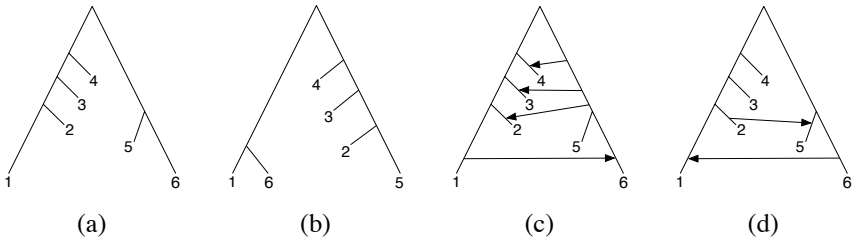


Figure 1.3 *(a) A species tree ST. (b) A tree GT of a horizontally transferred gene. Both networks in (c) and (d) are formed based on ST, and both induce GT. However, even though the actual HGT scenarios that took place are described by the network in (c), the HGT Reconstruction Problem seeks the solution in (d).*

in Figure 1.3(d) has fewer HGT events, yet induces GT. In this case, the solution in Figure 1.3(d) is the one sought by the HGT Reconstruction Problem. Although the scenarios depicted in Figure 1.3(c) and Figure 1.3(d) are very different, inferring the one in Fig 1.3(c) as the correct solution, in the absence of any additional biological knowledge about the organisms, would be rather arbitrary. Hence, based on the species and gene tree topologies, solving the HGT Reconstruction Problem offers the "best" solution. Another serious problem that impacts the identifiability of reticulate evolution is that of extinction and incomplete taxon sampling. Moret *et al*. (2004) illustrated some of the scenarios that lead to non-identifiability of reticulation events from a set of gene trees.

Hallett & Lagergren (2001) gave an efficient algorithm for solving the HGT Reconstruction Problem; however, their algorithm handles limited special cases of the problem in which the number of HGT events is very small, and the number of times a gene is transferred is very low (also, their tool handles only binary trees; Addario-Berry *et al*. (2003)). Nakhleh *et al*. (2004) gave efficient algorithms for solving the Hybrid Speciation Reconstruction Problem, but for constrained phylogenetic networks, referred to as *gt-networks*; further, they handled only binary trees. Nakhleh *et al*. (2005) have recently introduced RIATA-HGT, which is the first method for solving the general case of the HGT Reconstruction Problem. The method can be easily modified to yield a heuristic for solving the Hybrid Speciation Reconstruction Problem. We now describe the method and its empirical performance.

RIATA-HGT: reconstructing HGT from gene trees

We describe the algorithm underlying RIATA-HGT in terms of a species tree and a gene tree. The core of RIATA-HGT is the divide-and-conquer algorithm ComputeHGT algorithm (outlined in Figure 1.4). The algorithm starts by computing the $MAST$, T', of the species tree ST and gene tree GT; tree T' forms the basis for detecting and reconstructing the HGT events (computing T' is done in Step 1 in Figure 1.4). The algorithm then decomposes the clade sets U_1 and U_2 (whose removal from

ST and GT, respectively, yields T') into maximal clades such that each maximal clade in one of the two sets is "matched" by a maximal clade on the same leaf set in the second set. The algorithm for this decomposition is outlined in Figure 1.5. The algorithm then recurses on each maximal clade and its matching maximal clade (Steps 5.c.(1) and 5.d.(5).(1) in Figure 1.4) to compute the HGT events whose recipients form sub-clades of those maximal clades. Finally, we add a single HGT event per each maximal clade to connect it to its "donor" in the ST; this is achieved through the calls to AddSingleHGT (Figure 1.6) in Steps 5.c.(2) and 5.d.(5).(3) in Figure 1.4.

PROCEDURE COMPUTEHGT(ST,GT)
Input: Species tree ST, and gene tree GT, both on the same set S of taxa.
Output: Computes the set Ξ of HGT events such that the pair (ST, Ξ) induces GT.

1. $T' = MAST(ST, GT)$;
2. **If** $T' = ST$ **then**
   (a) **Return**;
3. $U_1 = ST - T'$; $U_2 = GT - T'$;
4. $V = \emptyset$;
5. **Foreach** $u_2 \in U_2$
   (a) $Decompose(U_1, u_2, T', V)$;
6. $U_2 = V$;
7. **While** $V \neq \emptyset$
   (a) Let u_2 be an element of V;
   (b) Let $u_1 \in U_1$ be such that $L(u_2) \subseteq L(u_1)$;
   (c) $Y = \{y \in U_2 : L(y) \cap L(u_1) \neq \emptyset\}$;
   (d) $Z = \{y | (L(y) - L(u_1)) : \ y \in Y\}$;
   (e) $V = V - Y$; $V = V \cup Z$;
   (f) $X = \{u_1 | L(y) : y \in Y\}$;
   (g) **Foreach** $y \in Y$
      i. Let $x \in X$ be such that $L(x) \cap L(y) \neq \emptyset$;
      ii. $ComputeHGT(x, y)$;
      iii. $AddSingleHGT(ST, GT, y, U_2, T')$;

Figure 1.4 *The main algorithm for detecting and reconstructing HGT events based on a pair of species tree and gene tree.*

Theoretically, RIATA-HGT may not compute the minimum-cardinality set of HGT events; Nakhleh *et al.* (2005) established the following properties of the method.

Theorem 1.1 *Given a species tree ST and a gene tree GT, the network N ob-*

PROCEDURE DECOMPOSE(U_1,u_2, T, U')
Input: Set U_1 of clades from ST, clade u_2 from GT, the backbone clade u_2, and U' which will contain the "refined" clades of u_2.
Output: Decompose u_2 so that no clade in U' has a leaf set that is the union of leaf sets of more than one clade in U_1.

1. **If** $\exists u_1 \in U_1$ such that $L(u_2) \subseteq L(u_1)$ **then**
   (a) $U' = U' \cup \{u_2\}$;
   (b) $B(u_2) = T$;
   (c) **Return** u_2;
2. **Else**
   (a) **If** $\exists u_1 \in U_1$ such that $r(u_2) = r(u_2|L(u_1))$
      i. $t = u_2|L(u_1)$;
      ii. $U' = U' \cup \{t\}$;
      iii. $B(t) = T$;
      iv. Let $X = u_2 - t$;
      v. **Foreach** $x \in X$
         A. $Decompose(U_1, x, t, U')$;
      vi. **Return** t;
   (b) **Else**
      i. Let $c_1, \ldots, c_k$ be the children of $r(u_2)$;
      ii. $x = Decompose(U_1, T_{c_1}, T, U')$;
      iii. **For** $i = 2$ to k
         A. $Decompose(U_1, T_{c_i}, x, U')$;
      iv. **Return** x;

Figure 1.5 *The algorithm for decomposing the clades in U_1 and U_2 such that for all $u_1 \in U_1$ and $u_2 \in U_2$ we have $L(u_1) \not\subset L(u_2)$.*

tained by running RIATA-HGT on (ST, GT) induces GT. Further, RIATA-HGT takes $O(n^4)$ time in the worst case, where n is the number of leaves in ST.

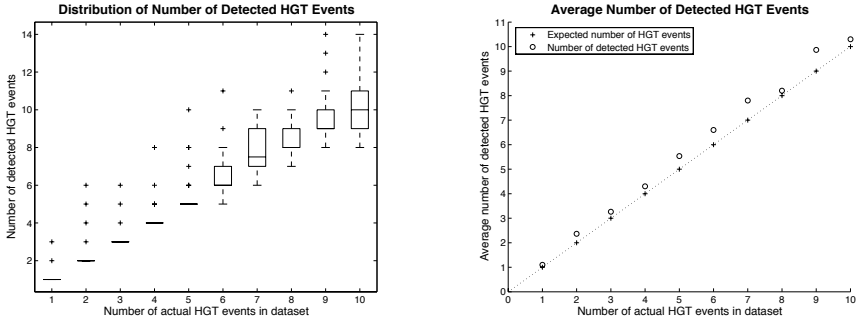
Moreover, experimental results show very good empirical performance on synthetic data, as illustrated in Figure 1.7. The whisker-and-box plot in Figure 1.7(a) shows the individual numbers of HGT events as predicted by RIATA-HGT versus the actual numbers. Figure 1.7(b) shows the average (of 30 runs) numbers of HGT events as predicted by RIATA-HGT versus the actual numbers (for full details of how the simulation studies were conducted and detailed analyses of the results, please refer to Nakhleh *et al.* (2005)). The plots demonstrate empirically the excellent performance of RIATA-HGT; it estimates the exact number of HGT events in a great majority of the cases, with very mild over- or under-estimation in the other cases. Over-estimation is an artifact of the heuristic nature of RIATA-HGT, whereas underestimation is an artifact of the parsimony criterion in the definition of the prob-

PROCEDURE ADDSINGLEHGT(ST, GT, u_2, U_2, T')
Input: Species tree ST, gene tree GT, clade u_2 of GT, set U_2 of clades of GT, and $MAST$ T' of ST and GT.
Output: Add to Ξ a single HGT event whose donor is determined in this procedure and whose recipient is clade u_2.

1. $Q = L(u_2) \cup L(B(u_2))$;
2. $T'' = GT|Q$; $p = lca_{T''}(L(u_2))$;
3. $tq = lca_{ST}(L(u_2))$; $te = inedge_{ST}(tq)$;
4. **If** p is a child of $r(T'')$ and $|L(B(u_2))| > 1$ **then**
   (a) $sq = lca_{ST}(L(B(u_2)))$;
   (b) $\Xi = \Xi \cup (sq \rightarrow te)$;
5. **Else**
   (a) $O = \bigcup_{\{p':p'=sibling_{T''}(p)\}} L(T_{p'})$;
   (b) $sq = lca_{ST}(O)$; $se = inedge_{ST}(sq)$;
   (c) $\Xi = \Xi \cup (se \rightarrow te)$;

Figure 1.6 *The algorithm for detecting and reconstructing the single HGT event in which clade u_2 is the recipient.*

lem (see the discussion above). RIATA-HGT was also applied to the bacterial gene



(a) (b)

Figure 1.7 *The results of RIATA-HGT on synthetic datasets. (a) A box-and-whisker plot for the predictions of HGT event numbers made by RIATA-HGT. (b) The averages of HGT event numbers estimated by RIATA-HGT vs. the actual number of HGT events. Each point is the average of 30 runs of RIATA-HGT.*

datasets reported in Lerat *et al*. (2003), and produced the results hypothesized by Lerat *et al*. In summary, RIATA-HGT performed very well on the synthetic datasets, as well as on the biological datasets.

1.5 The Coalescent and Reticulate Evolution

1.5.1 The coalescent and lineage sorting in ancestral populations

Intra-species events (*i.e.,* gene duplication and loss) occur because of random contribution of each individual to the next generation. Some fail to have offsprings (gene loss) while some happen to have multiple offsprings (gene duplication). This means a number of duplication and loss events occur every generation. In population genetics, this process was first modeled by R. A. Fisher and S. Wright, in which each gene of the population at a particular generation is chosen independently from the gene pool of the previous generation, regardless of whether the genes are in the same individual or in different individuals.

Under the Wright-Fisher model, "the coalescent" considers the process backward in time (Kingman (1982); Hudson (1983b); Tajima (1983)). That is, the ancestral lineages of genes of interest are traced from offsprings to parents. A coalescent event occurs when two (or sometimes more) genes are originated from the same parent, which is called the most recent common ancestor (MRCA) of the two genes. This event corresponds to gene duplication when the process is considered forward in time. Gene loss events can be ignored in the coalescent process, because we are not interested in the lineages that do not exist at present.

The basic process can be treated as follows. Consider a pair of genes at time τ_1 in a random mating haploid population. The population size at time τ is denoted by $N(\tau)$. The probability that the pair are from the same parental gene at the previous generation (time τ_1+1) is $1/N(\tau_1+1)$. Therefore, starting at τ_1, the probability that the coalescence between the pair occurs at τ_2 is given by

$$Prob(\tau_2) = \frac{1}{N(\tau_2)} \sum_{\tau=\tau_1+1}^{\tau_2-1} \left(\frac{1}{N(\tau)}\right). \tag{1.1}$$

When $N(\tau)$ is constant, the probability density distribution (pdf) of the coalescent time (*i.e.,* $t = \tau_2 - \tau_1$) is given by a geometric distribution, and can be approximated by an exponential distribution for a large N:

$$Prob(t) = \frac{1}{N} e^{-t/N}. \tag{1.2}$$

The coalescent process is usually ignored in phylogenetic analysis, but has a significant effect (causing lineage sorting) when closely related species are considered (Hudson (1983a); Takahata (1989); Rosenberg (2002)). The situation of Figure 1.1(b) is reconsidered under the framework of the coalescent in Figure 1.8. Here, it is assumed that species A and B split $T_1 = 5$ generations ago, and the ancestral species of A and B and species C split $T_2 = 19$ generation ago. The ancestral lineage of a gene from species A and that from B meet in their ancestral population at time $\tau = 6$, and they coalesce at $\tau = 35$, which predates T_2, the speciation time between (A, B) and C. The ancestral lineage of B enters in the ancestral population of the three species at time $\tau = 20$, and first coalesces with the lineage of C. Therefore,

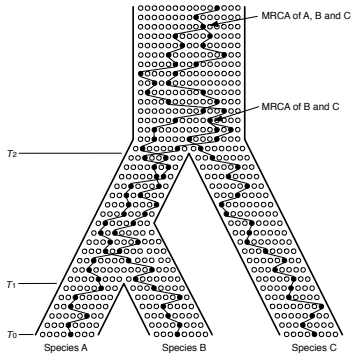


Figure 1.8 *An illustration of the coalescent process in a three species model with discrete generations. The process is considered backward in time from present, T_0, to past. Circles represent haploid individuals. We are interested in the gene tree of the three genes (haploids) from the three species. Their ancestral lineages are represented by closed circles connected by lines. A coalescent event occurs when a pair of lineages happen to share a single parental gene (haploid).*

the gene tree is represented by $A(BC)$ while the species tree is $(AB)C$. That is, the gene tree and species tree are "incongruent". Under the model in Figure 1.8, the probability that the gene tree is congruent with the species tree is 0.85, which is one minus the product of the probability that the ancestral lineages of A and B do not coalesce between $\tau = 6$ and $\tau = 9$, and the probability that the first coalescence in the ancestral population of the three species occur between (A and C) or (B and C). The former probability is $\frac{14}{15}\frac{12}{13}\frac{11}{12}...\frac{7}{8}\frac{7}{8} = 0.22$ and the latter is $\frac{2}{3}$.

Under the three-species model (Figure 1.8), there are three possible types of gene tree, $(AB)C$, $(AC)B$ and $A(BC)$. Let $Prob$[(AB)C], $Prob$[(AC)B] and $Prob$[A(BC)] be the probabilities of the three types of gene tree. These three probabilities are simply expressed with a continuous time approximation when all populations have equal and constant population sizes, N, where N is large:

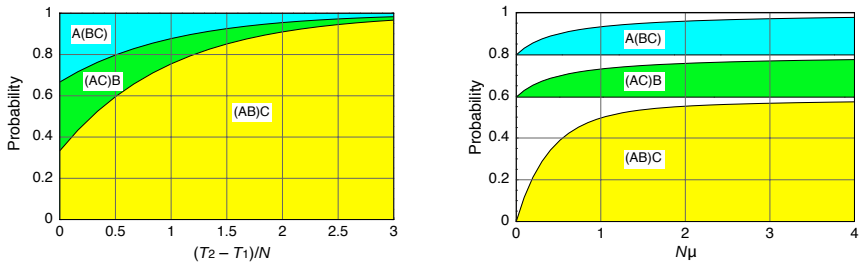
$$Prob[\text{(AB)C}] = 1 - \frac{2}{3}e^{-(T_2-T_1)/N}, \tag{1.3}$$

and

$$Prob[\text{(AC)B}] = Prob[\text{A(BC)}] = \frac{1}{3}e^{-(T_2-T_1)/N}. \tag{1.4}$$

Figure 1.9(a) shows the three probabilities as functions of $(T_2 - T_1)/N$.

An interesting application of this three species problem is in hominoids; A: human, B: chimpanzee and C: gorilla. It is believed that the species three is $(AB)C$. Chen & Li (2001) investigated DNA sequences from 88 autosomal intergenic regions, and the gene tree is estimated for each region. They found that 36 regions support the species tree, $(AB)C$, while 10 estimated trees are $(AC)B$ and 6 are $A(BC)$. No resolution is obtained for the remaining 36 regions (see below). It is possible to estimate the time

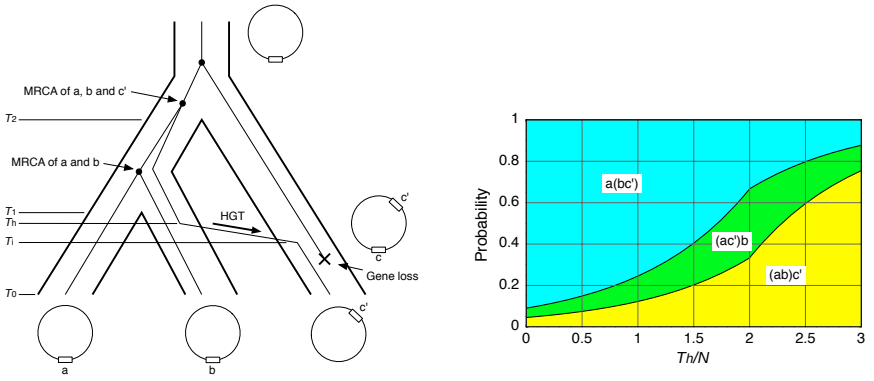


(a) (b)

Figure 1.9 *(a) The probabilities of the three types of gene tree, (AB)C, (AC)B, and A(BC), as functions of* $(T_2 - T_1)/N$. *(b) The probabilities that the gene tree is resolved from DNA sequence data. The probabilities are given functions of the mutation rate for the three types of tree, (AB)C, (AC)B, and A(BC), when* $(T_2 - T_1)/N = 0.5$. *The white regions represent the probabilities that the gene tree is not resolved.*

between two speciation events, $T_2 - T_1$, assuming all populations have equal and constant diploid population sizes, N (Wu (1991)). Since 36 out of 52 gene trees are congruent with the species tree, $T_2 - T_1$ is estimated to be $-\ln[(3/2)(36/52)] = 0.77$ times $2N$ generations. It should be noted that $2N$ is used for the coalescent time scale instead of N because hominoids are diploids. If we assume N to be $5 \times 10^4 - 1 \times 10^5$ (Takahata *et al.* (1995); Takahata & Satta (1997)), the time between two speciation events is $7.7 - 15.5 \times 10^4$ generations, which is roughly $1 - 3$ million years assuming a generation time of $15 - 20$ years.

It is important to notice that the estimation of the gene tree from DNA sequence data is based on the nucleotide differences between sequences, and that the gene tree is sometimes unresolved. One of the reasons for that is a lack of nucleotide differences such that DNA sequence data are not informative enough to resolve the gene tree. This possibility strongly depends on the mutation rate. Let μ be the mutation rate per region per generation, and consider the effect of mutation on the estimation of the gene tree. We consider the simplest model of mutations on DNA sequences, the infinite site model (Kimura (1969)), in which mutation rate per site is so small that no multiple mutations at a single site are allowed. Consider a gene tree, $(AB)C$, and suppose that we have a reasonable outgroup sequence such that we know the sequence of the MRCA of the three sequences. It is obvious that mutations on the internal branch between the MRCA of the three and the MRCA of A and B are informative. If at least one mutation occurred on this branch, the gene tree can be resolved from the DNA sequence alignment. This effect is investigated by assuming that the number of mutations on a branch with length t follows a Poisson distribution with mean μt. Figure 1.9(b) shows the probability that the gene tree is resolved; $T_2 - T_1 = 0.5N$ generations is assumed so that the probability that the gene tree is $(AB)C$ is about 0.6. As expected, as the mutation rate increases, the probability that the gene tree is resolved from the sequence alignment increases, and this probability



(a) (b)

Figure 1.10 *(a) A three species model with a HGT event. A demonstration that a congruent tree could be observed even with HGT. (b) The probabilities of the three types of gene tree, (ab)c', (ac')b, and a(bc'), as functions of* T_h/N. $T_1 = 2N$ *and* $T_2 = 3N$ *are assumed.*

exceeds 90% when $N\mu > 1.52$. Similar results are obtained for the other two types of trees, $(AC)B$ and $A(BC)$, that appears with probability 0.2 for each (see also Figure 1.9(b)).

1.5.2 Gene trees, species trees and reticulate evolution

In the previous section, we have shown that the gene tree is not always identical to the species tree. With keeping this in mind, let us consider the effect of horizontal gene transfer (HGT) on gene tree under the framework of the coalescent.

The application of the coalescent theory to bacteria is straightforward. Bacterial evolution is better described by the Moran model rather than the Wright-Fisher model because bacteria do not fit a discrete generation model. Suppose that each haploid individual in a bacterial population with size N has a lifespan that follows an exponential distribution with mean l. When an individual dies, another individual randomly chosen from the population replaces it to keep the population size constant. In other words, one of the $N - 1$ alive lineages is duplicated to replace the dead one. Under the Moran model, the ancestral lineages of individuals of interest can be traced backward in time, and the coalescent time between a pair of individuals follows an exponential distribution with mean $lN/2$ (Ewens (1979); Rosenberg (2005)). This means that one half of the mean lifetime in the Moran model corresponds to one generation in the Wright-Fisher model.

It may usually be thought that HGT can be detected when the gene tree and species tree are incongruent (see Section 1.4). However, the situation is complicated when lineage sorting is also involved. Consider a model with three species, A, B, and C, in which an HGT event occurs from species B to C. Suppose the ancient circular

genome has a single copy of a gene as illustrated in Figure 1.10(a). Let a, b and c be the focal orthologous genes in the three species, respectively. At time T_h, a gene escaped from species B and was inserted in a genome in species C at T_i, which is denoted by c'. Following the HGT event, c was physically deleted from the genome, so that each of the three species currently has a single copy of the focal gene.

If there is no lineage sorting, the gene tree should be $a(bc')$. Since this tree is incongruent with the species tree, $(AB)C$, we could consider it as an evidence for HGT. However, as demonstrated in Section 1.2, lineage sorting could also produce the incongruence between the gene tree and species tree without HGT. It is also important to note that lineage sorting, coupled with HGT, could produce congruent gene tree, as illustrated in Figure 1.10(a). Although b and c' have more chance to coalesce first, the probability that the first coalescence occurs between a and b or between a and c' may not be negligible especially when $T_1 - T_h$ is short.

The probabilities of the three types of gene tree can be formulated under this tri-species model with HGT as illustrated in Figure 1.10(a). Here, T_h could exceed T_1, in such a case it can be considered that HGT occurred before the speciation between A and B. Assuming that all populations have equal and constant population sizes, N, the three probability can be obtained modifying (1.3) and (1.4):

$$Prob[\text{(AB)C}] = \begin{cases} \frac{1}{3}e^{-(T_1-T_h)/N} & \text{if } T_h \leq T_1 \\ 1 - \frac{2}{3}e^{-(T_h-T_1)/N} & \text{if } T_h > T_1 \end{cases}, \tag{1.5}$$

$$Prob[\text{(AC)B}] = \begin{cases} \frac{1}{3}e^{-(T_1-T_h)/N} & \text{if } T_h \leq T_1 \\ \frac{1}{3}e^{-(T_h-T_1)/N} & \text{if } T_h > T_1 \end{cases}, \tag{1.6}$$

and

$$Prob[\text{A(BC)}] = \begin{cases} 1 - \frac{2}{3}e^{-(T_1-T_h)/N} & \text{if } T_h \leq T_1 \\ \frac{1}{3}e^{-(T_h-T_1)/N} & \text{if } T_h > T_1 \end{cases}. \tag{1.7}$$

Figure 1.10(b) shows the three probability assuming $T_1 = 2N$ and $T_2 = 3N$.

Thus, lineage sorting due to the coalescent process works as a noise for detecting and reconstructing HGT based on gene tree, sometimes mimicking the evidence for HGT and sometimes creating a false positive evidence for HGT. Therefore, to distinguish HGT and lineage sorting, statistics based on the theory introduced in this chapter is needed. We only considered very simple cases with three species here, but it is straightforward to extend the theory to more complicated models.

1.6 Summary

In this chapter, we have reconsidered the gene tree species tree problem in the context of reticulate evolution. In particular, we discussed gene tree incongruence due to reticulate evolution and presented our recent heuristic, RIATA-HGT, for resolving this type of incongruence. Further, we have addressed extensions of the coalescent model to incorporate non-treelike evolutionary events, such as horizontal gene transfer. Gene tree incongruence is both an obstacle impeding accurate phylogeny

reconstruction and a tool for detecting and reconstructing evolutionary events such as HGT and hybrid speciation. Future directions for further research include:

1. Testing the performance of existing methods for resolving gene tree incongruence in the context of intra- and inter-species evolutionary events.
2. Developing and testing accurate and fast methods for reconstructing phylogenetic networks from gene trees under the conditions of incomplete taxon sampling and missing orthologs.
3. Extending our initial progress on the coalescent model beyond three species and to incorporate hybrid speciation and meiotic recombination.

Bibliography

Addario-Berry, L., Hallett, M.T., & Lagergren, J. 2003. Towards identifying lateral gene transfer events. *Pages 279–290 of: Proc. 8th Pacific Symp. on Biocomputing (PSB03).*

Arvestad, L., Berglund, A.-C., Lagergren, J., & Sennblad, B. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Pages i7–i15 of: Proc. 11th Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB03).* Bioinformatics, vol. 19.

Arvestad, Lars, Berglund, Ann-Charlotte, Lagergren, Jens, & Sennblad, Bengt. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *Pages 326–335 of: Proc. 8th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB04).*

Charleston, M. A. 2000. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences.*

Chen, Feng-Chi, & Li, Wen-Hsiung. 2001. Genomic Divergences between Humans and Other Hominoids and the Effective Population Size of the Common Ancestor of Humans and Chimpanzees. *Am. J. Hum. Genet.*, **68**, 444–456.

Chen, K., Durand, D., & Farach-Colton, M. 2000. Notung: A Program for Dating Gene Duplications and Optimizing Gene Family Trees. *Journal of Computational Biology*.

Daubin, V., Moran, N.A., & Ochman, H. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science*, **301**, 829–832.

Doolittle, W.F. 1999a. Lateral genomics. *Trends in Biochemical Sciences*, **24**(12), M5–M8.

Doolittle, W.F. 1999b. Phylogenetic classification and the universal tree. *Science*, **284**, 2124–2129.

Durand, D., Halldorsson, B., & Vernot, B. 2005. A Hybrid Micro-Macroevolutionary Approach to Gene Tree Reconstruction. *Pages 250–264 of: recomb05.*

Enright, M.C., Robinson, D.A., Randle, G., Feil, E.J., Grundmann, H., & Spratt, B.G. 2002. The evolutionary history of methicillin-resistant Staphylococcus aureus (MRSA). *Proc. Natl. Acad. Sci. USA*, **99**(11), 7687–7692.

et al., M. McClilland. 2004. Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid.

Nature Genetics, **36**(12), 1268–1274.

et al., R.A. Welch. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc. Nat'l Acad. Sci., USA*, **99**(26), 17020–17024.

Eulenstein, O. 1997. A linear time algorithm for tree mapping. *Arbeitspapire der GMD*, **1046**.

Eulenstein, O., Mirkin, B., & Vingron, M. 1996. Comparison of an annotatng duplication, tree mapping, and copying as methods to compare gene trees within species trees. *Mathematical Hierarchies and Biology, DIMACS Series in Discrete mathematics and Theoretical Computer Science*, **37**, 71–93.

Eulenstein, O., Mirkin, B., & Vingron, M. 1998. Duplication-based measures of difference between gene and species trees. *J. Comput. Biol.*, **5**, 135–148.

Ewens, Warren J. 1979. *Mathematical Population Genetics*. Berlin: Springer-Verlag.

Fellows, M.R., Hallett, M.T., Korostensky, C., & Stege, U. 1998. Analogs & duals of the MAST problem for sequences & trees. *Pages 103–114 of: Proc. Eur. Symp. Algs. ESA98.* in *LNCS 1461*.

Hallett, M.T., & Lagergren, J. 2000. New algorithms for the duplication-loss model. *Pages 138–146 of: Proc. 4th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB00).* New York: ACM Press.

Hallett, M.T., & Lagergren, J. 2001. Efficient algorithms for lateral gene transfer problems. *Pages 149–156 of: Proc. 5th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB01).* New York: ACM Press.

Hao, W., & Golding, G.B. 2004. Patterns of Bacterial Gene Movement. *Mol. Biol. Evol.*, **21**(7), 1294–1307.

Hein, J. 1990. Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosciences*, **98**, 185–200.

Hillis, D.M., & Huelsenbeck, J.P. 1994. To tree the truth: Biological and numerical simulations of phylogeny. *Pages 55–67 of:* Fambrough, D.M. (ed), *Molecular Evolution of Physiological Processes*. Rockefeller University Press.

Hillis, D.M., & Huelsenbeck, J.P. 1995. Assessing molecular phylogenies. *Science*, **267**, 255–256.

Hillis, D.M., Bull, J.J., White, M.E., Badgett, M.R., & Molineux, I.J. 1993. Experimental approaches to phylogenetic analysis. *Syst. Biol.*, **42**, 90–92.

Ho, M.-W. 2002. *Recent Evidence confirms risks of horizontal gene transfer*. http://www.i-sis.org.uk/FSAopenmeeting.php.

Hudson, R. R. 1983a. Testing the Constant-Rate Neutral Allele Model with Protein Sequence Data. *Evolution*, **37**, 203–217.

Hudson, R.R. 1983b. Properties of the neutral allele model with intergenic recombination. *Theor. Popul. Biol.*, **23**, 183–201.

Kimura, Motoo. 1969. The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population due to Steady Flux of Mutations. *Genetics*, **61**, 893–903.

Kingman, J. F. C. 1982. The Coalescent. *Stochast. Proc. Appl.*, **13**, 235–248.

Kurland, C.G., Canback, B., & Berg, O.G. 2003. Horizontal gene transfer: A critical view. *Proc. Nat'l Acad. Sci., USA*, **100**(17), 9658–9662.

Lerat, E., Daubin, V., & Moran, N.A. 2003. From Gene Trees to Organismal Phylogeny in Prokaryotes: The case of the γ-Proteobacteria. *PLoS Biology*, **1**(1), 1–9.

Lewis, Ricki. 1995. *The rise of antibiotic-resistant infections*. http://www.fda.gov/fdac/features/795_antibio.html.

Ma, B., Li, M., & Zhang, L. 1998. On reconstructing species trees from gene trees in terms of duplications and losses. *Pages 182–191 of: Proc. 2nd Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB98)*.

Ma, B., Li, M., & Zhang, L. 2000. From gene trees to species trees. *SIAM Journal on Computation*.

Maddison, W.P. 1995. Phylogenetic histories within and among species. *Experimental and molecular approaches to plant biosystematics. Monographs in systematics*, **53**, 273–287. Missouri Botanical Garden, St. Louis.

Maddison, W.P. 1997. Gene Trees in Species Trees. *Systematic Biology*, **46**(3), 523–536.

Mirkin, B., Muchnik, I., & Smith, T. 1995. A biologically consistent model for comparing molecular phylogenies. *J. Comput. Biol.*, **2**(4), 493–507.

Moore, WS. 1995. Inferring phylogenies from mtDNA variation: mitochondrial gene trees versus nuclear gene trees. *Evolution*, **49**, 718–726.

Moret, B.M.E., Roshan, U., & Warnow, T. 2002. Sequence length requirements for phylogenetic methods. *Pages 343–356 of: Proc. 2nd Int'l Workshop Algorithms in Bioinformatics (WABI02)*. Lecture Notes in Computer Science, vol. 2452.

Moret, B.M.E., Nakhleh, L., Warnow, T., Linder, C.R., Tholse, A., Padolina, A., Sun, J., & Timme, R. 2004. Phylogenetic networks: modeling, reconstructibility, and accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **1**(1), 13–23.

Nakamura, Y., Itoh, T., Matsuda, H., & Gojobori, T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature Genetics*, **36**(7), 760–766.

Nakhleh, L., Roshan, U., John, K. St., Sun, J., & Warnow, T. 2001a. Designing Fast Converging Phylogenetic Methods. *Bioinformatics*, **17**(90001), S190–S198. ISMB01 Conference.

Nakhleh, L., Roshan, U., John, K. St., Sun, J., & Warnow, T. 2001b. The Performance of Phylogenetic Methods on Trees of Bounded Diameter. *Pages 214–226 of:* Gascuel, O., & Moret, B.M.E. (eds), *Proc. 1st Int'l Workshop Algorithms in Bioinformatics (WABI01)*. Lecture Notes in Computer Science, vol. 2149.

Nakhleh, L., Moret, B.M.E., Roshan, U., John, K. St., Sun, J., & Warnow, T. 2002. The Accuracy of Phylogenetic Methods for Large Datasets. *Pages 211–222 of: Proc. 7th Pacific Symp. on Biocomputing (PSB02)*, vol. 7.

Nakhleh, L., Warnow, T., & Linder, C.R. 2004. Reconstructing reticulate evolution in species–theory and practice. *Pages 337–346 of: Proc. 8th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB04)*.

Nakhleh, L., Ruths, D., & Wang, L.S. 2005. RIATA-HGT: A Fast and accurate heuristic for reconstrucing horizontal gene transfer. *In: Proc. 11th Int'l Conf. Computing and Combinatorics (COCOON05)*.

Nichols, Richard. 2001. Gene trees and species trees are not the same. *Trends in Ecology and Evolution*, **16**(7).

Ochman, H., Lawrence, J.G., & Groisman, E.A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**(6784), 299–304.

Page, R. 1998. GeneTree: Comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, **14**(9), 819–820.

Page, R., & Charleston, M.A. 1997a. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Mol. Phyl. Evol.*, **7**, 231–240.

Page, R. D. M., & Charleston, M. A. 1997b. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution*, **7**, 231–240.

Page, R. D. M., & Charleston, M. A. 1998. Trees within trees: phylogeny and historical associations. *Trends in Ecology and Evolution*.

Page, R.D.M. 1990. Component analysis: a valiant failure? *Cladistics*, **6**, 119–136.

Page, R.D.M. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, **43**, 58–77.

Paulsen, I.T., Banerjei, L., Myers, G.S.A., & *et al.*, K.E. Nelson. 2003. Role of Mobile DNA in the Evolution of Vacomycin-resistant Enterococcus faecalis. *Science*, **299**(5615), 2071–2074.

Rosenberg, N. 2002. The probability of topological concordance of gene trees and species tree. *Theoretical Population Biology*, **61**, 225–247.

Rosenberg, N. A. 2005. Gene genealogies. *Chap. 15 of:* Fox, C.W., & Wolf, J. B. (eds), *Evolutionary Genetics: Concepts and Case Studies*. Oxford Univ. Press University Press.

Ruvolo, Maryellen. 1997. Molecular Phylogeny of the Hominoids: Inferences from Multiple Independent DNA Sequence Data Sets. *Molecular Biology and Evolution*, **14**(3).

Stege, U. 1999a. Gene trees and species trees: the gene-duplication problem is fixed-parameter tractable. *In: Proc. 6th Workshop Algorithms and Data Structures (WADS99)*. Lecture Notes in Computer Science, vol. 1663. Springer-Verlag.

Stege, U. 1999b. Gene trees and species trees: The gene-duplication problem is fixed-parameter tractable. *In: Proceedings of the 6th International workshop on Algorithms and Data Structures (WADS'99)*.

Tajima, Fumio. 1983. Evolutionary Relationship of DNA Sequences in Finite Populations. *Genetics*, **105**, 437–460.

Takahata, N., & Satta, Y. 1997. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proc. Natl. Acad. Sci. USA*, **94**, 4811–4815.

Takahata, Naoyuki. 1989. Gene Genealogy in Three Related Populations: Consistency Probability Between Gene and Population Trees. *Genetics*, **122**, 957–966.

Takahata, Naoyuki, Satta, Yoko, & Klein, Jan. 1995. Divergence Time and Population Size in the Lineage Leading to Modern Humans. *Theor. Pop. Biol.*, **48**, 198–221.

Wu, Chung-I. 1991. Inferences of Species Phylogeny in Relation to Segregation of Ancient Polymorphisms. *Genetics*, **127**, 429–435.

Zhang, L. 1997. On a Mirkin-Muchnik-Smith Conjecture for Comparing Molecular Phylogenies. *Journal of Computational Biology*, **4**(2), 177–187.

Zmasek, C. M., & Eddy, S. R. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, **17**(9), 821–828.