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
Graduate University for Advanced Studies
Hayama, Kanagawa 240-0193, Japan
innan_hideki@soken.ac.jp

1.1 Introduction

The availability of whole-genome sequence data from multiple organisms has provided a rich resource for investigating several biological, medical and pharmaceutical problems and applications. Yet, along with the insights and promises these data are providing, genome-wide data have given rise to more complex problems and challenged *traditional* biological paradigms. One of these paradigms is the inference of a *species phylogeny*. Traditionally, a biologist would infer the phylogeny, or evolutionary history, of a set of species by obtaining the molecular sequence of a single locus, or gene, in these species, inferring the phylogeny of this locus, and taking it to be an accurate representation of the species pattern of divergence. While this approach may work for several groups of organisms, particularly when taking extra caution in selecting the locus, the availability of sequence data for multiple loci from a variety of organisms and populations has highlighted the deficiencies and inaccuracies of this traditional approach. Different loci in a group of organisms may have different gene tree topologies. In this case, there is no single gene tree topology to declare as the species tree. Further, in the presence of *reticulate* evolutionary events, such as horizontal gene transfer, the species phylogeny may not be a tree; instead, a *network* of relationships is the more appropriate model.

In a seminal paper, Maddison (1997) discussed the issue of species/gene tree incongruence, the implications it has on the inference of a species tree, and the processes that can cause such incongruence and whose explicit modeling is necessary for accurate inferences. The three main processes that Maddison discussed were *lineage sorting*, *gene duplication and loss*, and *reticulate evolution*.

Lineage sorting occurs because of random contribution of genetic material from each individual in a population to the next generation. Some fail to have offsprings while some happen to have multiple offsprings. 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 "merge" at the same parent, which is called the most recent common ancestor (MRCA) of the two genes. In certain cases, two genes coalesce at a branch in the species tree that is deeper than their MRCA, and when this happens, it may be that coalescence patterns result in trees that do not reflect that divergence patterns of the species. Evidence of extensive lineage sorting has been reported in several groups of organisms; e.g., (Rokas *et al.* (2003); Syring *et al.* (2005); Pollard *et al.* (2006); Than *et al.* (2008a); Kuo *et al.* (2008)).

Gene duplication is considered a major mechanism of evolution, particularly of generating new genes and biological functions (Ohno (1970); Graur & Li (2000)). Duplication events result in multiple gene copies, and when those are transmitted to descendant organisms, complex gene genealogies result. As some of these genes may go extinct (Olson (1999)), inferring the gene tree from the only copies present in the organisms result in a topology that may disagree with that of the species tree. A similar effect can be obtained as an artifact of sampling some, but not all, of the gene copies in an organism's genome.

The third process discussed by Maddison (1997) is reticulate evolution. For example, evidence shows that bacteria may obtain a large proportion of their genetic diversity through the acquisition of sequences from distantly related organisms, via horizontal gene transfer (HGT) (Ochman *et al.* (2000); Doolittle (1999b,a); Kurland *et al.* (2003); Hao & Golding (2004); Nakamura *et al.* (2004)). Furthermore, more evidence of widespread HGT in plants is emerging recently (Bergthorsson *et al.* (2003, 2004); Mower *et al.* (2004)). Interspecific recombination is believed to be ubiquitous among viruses (Posada *et al.* (2002); Posada & Crandall (2002)), and hybrid speciation is a major evolutionary mechanisms in plants, and groups of fish and frogs (Ellstrand *et al.* (1996); Rieseberg & Carney (1998); Rieseberg *et al.* (2000); Linder & Rieseberg (2004); Mallet (2005); Noor & Feder (2006); Mallet (2007)).

There is a major difference between lineage sorting and gene duplication/loss on the one hand and reticulate evolutionary events on the other, in terms of the reconciliation outcome. While gene trees may disagree with each other, as well as with the species phylogenies due to lineage sorting or gene duplication/loss events, in this case, their reconciliation yields a tree topology, with the deep coalescences, duplications, and losses taking place within the species tree branches. However, when horizontal gene transfer or hybrid speciation occur, the evolutionary history of the genomes can no longer be modeled by a tree; instead, a *phylogenetic network* is the more appropriate model.

Recognizing the presence of these processes, and incorporating them into the computational methods for inferring accurate evolutionary histories, will have significant implications on reconstructing accurate evolutionary histories of genomes and better understanding of their diversification. 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 by reconciling incongruent gene trees.

In this chapter, we review the processes that cause gene tree incongruence, issues that must be accounted for when dealing with these processes, and methods for reconciling gene trees and inferring species phylogenies despite gene tree incongruence. The rest of the chapter is organized as follows. In Section 1.2 we discuss in more detail the processes that result in discord. In Section 1.3, we discuss approaches for reconciling gene trees when incongruence is solely due to lineage sorting. In Section 1.4, we discuss incongruence due to gene duplication and loss, and review some of the methods for reconciliation under this scenario. In Section 1.5, we describe the *phylogenetic network* model and discuss the problem of reconciling gene trees into species networks, assuming that incongruence is due to reticulate evolutionary events. In Section 1.6 we discuss preliminary attempts at establishing unified frameworks for distinguishing among the various processes. Such frameworks are crucial to accurate reconstruction of species phylogenies, since in general the cause of incongruence may not be known, and assuming one of the three processes arbitrarily may result in wrong reconciliations. We conclude the chapter in Section 1.7.

1.2 Gene Tree Incongruence

A **gene tree** is a model of how a gene evolves through not only nucleotide substitution, but also other mechanisms that act on a larger scale, such as duplication, loss, and horizontal gene transfer. 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 (statistical error). Various studies show the effects of statistical error 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)). Statistical errors 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 the three aforementioned processes, on the other hand, is a reflection of true biological events.

We illustrate in Figure 1.1(a) the effect of lineage sorting, gene duplication and loss, and reticulate evolution on gene tree incongruence. The species phylogeny is represented by the shaded "tubes"; it has B and C as sister taxa whose most recent common ancestor (MRCA) is a sister taxon of A, and the MRCA of all three taxa is a sister taxon of D. In the case of hybrid speciation (the scenario in Figure 1.1(d)), the MRCA of taxa B and C is a sister taxon of both A and D, since it is the outcome of hybridization. Each of the four scenarios shows the trees of two genes evolving within the branches of the species phylogeny. For clarity, the topologies of the two gene trees are shown separately in Figure 1.2. We elaborate more on how such a scenario arise in Section 1.3.

Figure 1.1(b) shows an evolutionary scenario involving only gene duplication and loss events that result in identical gene tree topologies to that of the lineage sorting scenario. In this scenario, the gene whose tree is depicted by GT_1 had a divergence pattern identical to that of the species. The second gene, on the other hand, underwent a duplication event prior to the splitting of D from the MRCA of the other three taxa, and then copies of the gene went extinct in the D lineage, the A lineage, and the lineage of the MRCA of B and C. This combination of duplication and loss events results in gene tree GT_2 whose topology disagrees with that of GT_1 as well as the species tree. The topologies of the two gene trees in this scenario are also identical to the ones shown in Figure 1.2.

As can be seen in Figures 1.1(a) and 1.1(b), even though the two gene trees are incongruent, they are reconciled within the branches of a species tree that is identical in both cases. In other words, the incongruence in these two scenarios does not necessitate deviating from a tree-like pattern of divergence at the species level. This stands in contrast to the two scenarios illustrated in Figures 1.1(c) and 1.1(d), where reticulate evolutionary events (in this case, horizontal gene transfer and hybrid speciation, respectively) do necessitate the adoption of a phylogenetic network as the more appropriate model of the evolutionary history of the genomes.

Views as to the extent of HGT in bacteria vary between the two extremes (Doolittle (1999b,a); Kurland *et al.* (2003); *et al.* (2002); Hao & Golding (2004); *et al.* (2004); Nakamura *et al.* (2004)). There is a big "ideological and rhetorical" gap between the researchers believing that HGT is so rampant, that a prokaryotic phylogenetic tree

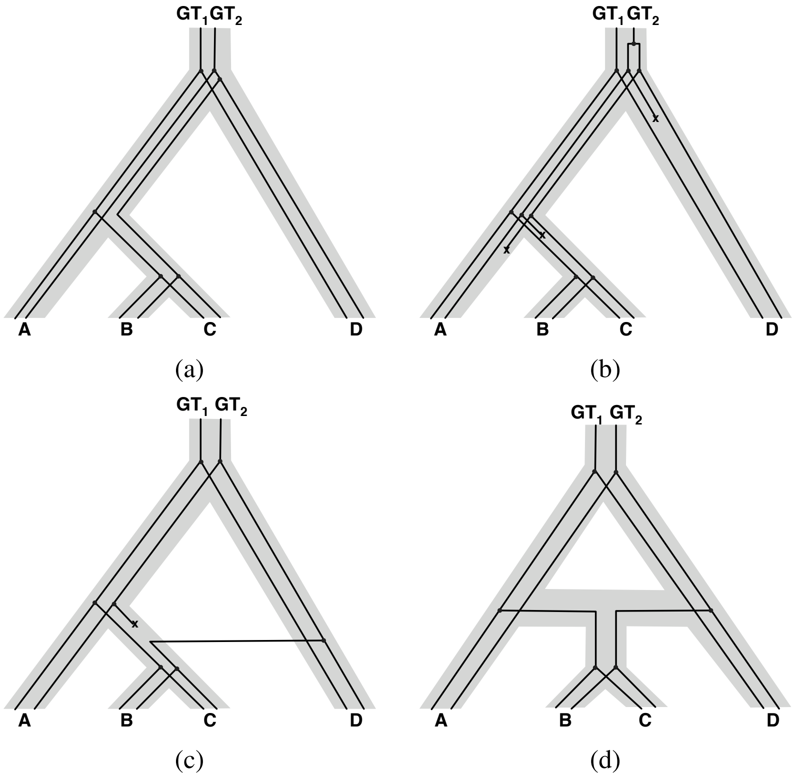

Figure 1.1 Four different evolutionary scenarios that result in gene tree incongruence. *(a) Gene tree GT_1 has a topology that is identical to that of the species phylogeny, whereas gene tree GT_2 in incongruent with both the species phylogeny and GT_1, due to lineage sorting. (b) Gene tree GT_1 has a topology that is identical to that of the species phylogeny, whereas gene tree GT_2 in incongruent with both the species phylogeny and GT_1, due to multiple duplication/loss events. (c) Gene tree GT_1 has a topology that is identical to that of the species phylogeny, whereas gene tree GT_2 in incongruent with both the species phylogeny and GT_1, due to horizontal gene transfer. (d) The species phylogeny is a network, since the clade (B, C) is a hybrid; the two gene trees GT_1 and GT_2 are incongruent due to hybrid speciation. The two different gene trees that arise from each of the four scenarios are shown in Figure 1.2.*

is useless, as opposed to those who believe HGT is mere “background noise” that does not affect the reconstructibility of a phylogenetic tree for bacterial genomes. Supporting arguments for these two views have been published. For example, the heterogeneity of genome composition between closely related strains (only 40% of the genes in common with three *E. coli* strains (Welch *et al.* (2002))) supports the former view, whereas the well-supported phylogeny reconstructed by Lerat *et al.* from about 100 “core” genes in γ-Proteobacteria (Lerat *et al.* (2003)) gives evidence in favor of the latter view. Nonetheless, regardless of the views and the accuracy of

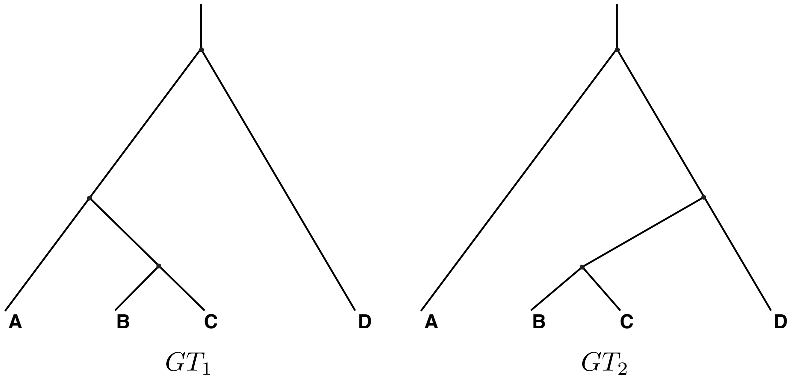

Figure 1.2 *The two gene trees GT_1 and GT_2 that arise in each of the four evolutionary scenarios depicted in Figure 1.1.*

the various analyses, the occurrence of HGT may result in networks, rather than trees, of evolutionary relationships of the genomes.

Three mechanisms of HGT are (1) *transformation*, which is the uptake of naked DNA from the environment (e.g., see Stewart & Sinigalliano (1991); Woegerbauer *et al.* (2002)); (2) *conjugation*, which is the transfer of DNA by direct physical interaction between a donor and a recipient, usually mediated by conjugal plasmids or conjugal transposons (e.g., see Nakamura *et al.* (2004)); and (3) *transduction*, which is the transfer of DNA by phage infection Brussow *et al.* (2004). Transducing phages have been observed in many bacteria including *Salmonella* (Schicklmaier & Schmieger (1995)), *Streptomyces* (Burke *et al.* (2001)), and *Listeria* (Hodgson (2000)). Transduction typically occurs among closely related bacteria, whereas conjugation may occur among more distant organisms.

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 GT_1), while other sites are horizontally transferred from another species (as in GT_2).

In the case of hybrid speciation, as illustrated in Figure 1.1(d), 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(d) shows a network with one hybrid speciation event, followed by speciation that results in the clade (B, D). Assuming no (incomplete) lineage sorting (i.e., that all alleles from B and C coalesce at the MRCA of these two taxa), each site evolves down exactly one of the two trees shown in Figures 1.2(a) and 1.2(b).

In the next three sections, we briefly address issues and methodologies for reconciling incongruence gene trees due to lineage sorting, gene duplication and loss, and reticulate evolutionary events, respectively.

1.3 Lineage Sorting

The basic coalescent 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 $\tau_1 + 1$) 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

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

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 1.3. 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 $\tau = 33$, 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, 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 1.3, the probability that the gene tree is congruent with the species tree is 0.863, which is one minus the product of the probability that the ancestral lineages of A and B do

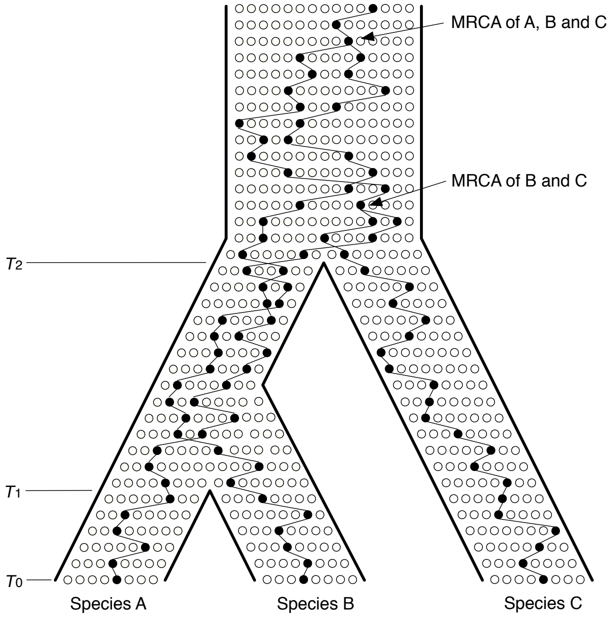

Figure 1.3 *An illustration of the coalescent process in a three species model with discrete generations. The process is considered backward in time from present, T_0, to past. Circles represent haploid individuals. We are interested in the gene tree of the three genes (haploids) from the three species. 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).*

not coalesce between $\tau = 6$ and $\tau = 9$, and the probability that the first coalescence in the ancestral population of the three species occurs between (A and C) or (B and C). The former probability is $\frac{15}{16}\frac{14}{15}\frac{12}{13}\frac{11}{12}\frac{10}{11}\frac{8}{9}\cdots(1-\frac{7}{8})^8 = 0.26$ and the latter is $\frac{2}{3}$.

Under the three-species model (Figure 1.3), 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

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

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 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.

A few methods have been introduced recently for analyzing gene trees, reconciling their incongruities, and inferring species trees despite these incongruities, when these incongruities are assumed to be caused solely by lineage sorting. Generally speaking, each of these methods follows one of two approaches: the *combined analysis* approach or the *separate analysis* approach; see Fig. 1.4. In the combined analysis aproach, the sequences from multiple loci are concatenated, and the resulting "supergene" data set is analyzed using traditional phylogenetic methods, such as maximum parsimony and maximum likelihood; e.g., (Rokas *et al.* (2003)). In the separate analysis approach, the sequence data from each locus is first analyzed individually, and a reconciliation of the gene trees is then sought. One way to reconcile the gene trees is by taking their majority consensus; e.g., (Kuo *et al.* (2008)). Another is the "democratic vote" method, which entails taking the tree topology occurring with the highest frequency among all gene trees as the species tree. Shortcomings of these methods based on the two approaches have been analyzed by various researchers (Degnan & Rosenberg (2006); Kubatko & Degnan (2007)). Recently, Bayesian methods following the separate analysis approach were developed (Edwards *et al.* (2007); Liu & Pearl (2007)). While these methods are accurate, they are very time consuming,

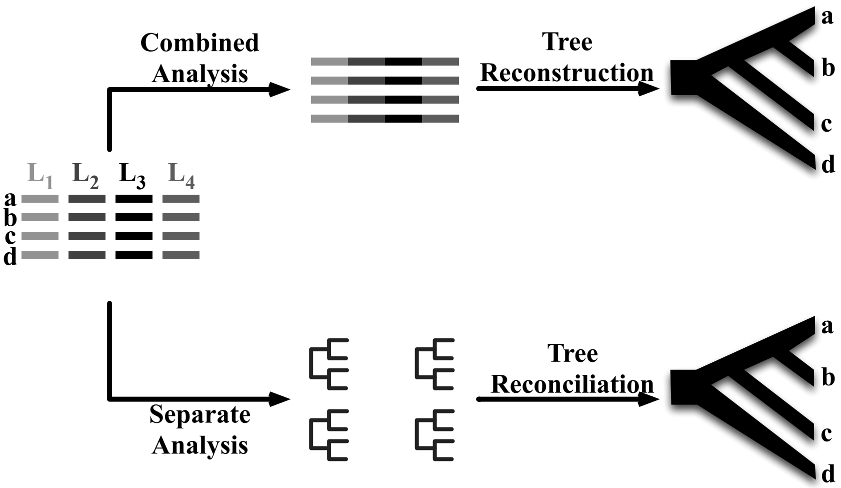

Figure 1.4 Approaches for inferring species trees. *In the combined analysis approach (top), gene sequences are concatenated, to create a "super gene," and a gene tree is inferred from this super gene. In the separate analysis approach (bottom), each gene is analyzed individually, and the analyses are then integrated in the form of gene tree reconciliation.*

taking hours and days even on moderate-size data sets, which limits their scalability (for example, the analysis of the yeast data set of Rokas *et al.* (2003) as conducted by the Bayesian approach of Edwards *et al.* (2007) took about 800 hours.) Than *et al.* (2008a) and Than & Nakhleh (2009) have recently introduced very efficient *integer linear programming* (ILP) formulations of the "minimize deep coalescences" approach, proposed by Maddison (1997) and Maddison & Knowles (2006). The resulting methods are very efficient, inferring accurate species trees in mere seconds, from data sets containing thousands of loci.

1.4 Gene Duplication and Loss

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 (1997b)). 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 (1997a); 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 (1997a)), 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 is 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)).

1.5 Reticulate Evolution

As described in Section 1.2, when events such as horizontal gene transfer or hybrid speciationoccur, 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 a phylogenetic network model that models reticulate evolution explicitly, and discuss approaches for reconstructing networks from gene trees.

1.5.1 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.

1.5.2 Evolutionary 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 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 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 v is a tree node; it is a* network edge *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(d) induces both trees shown in Figure 1.2(a) and Figure 1.2(b).

1.5.3 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, 2005b)). 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. A major issue 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.* (2005b) introduced RIATA-HGT for solving the general case of the HGT Reconstruction Problem, and later extended to handle non-binary trees as well as identify multiple minimal solutions by Than & Nakhleh (2008). The method is implemented in the PhyloNet software package by Than *et al.* (2008c). Other methods for reconciling species and gene trees, assuming reticulate evolution, include T-REX (Makarenkov (2001)), HorizStory (MacLeod *et al.* (2005)), and EEEP (Beiko & Hamilton (2006)).

Recently, Nakhleh and colleagues extended the maximum parsimony and maximum likelihood criteria to the domain of phylogenetic network reconstruction and evaluation (Nakhleh *et al.* (2005a); Jin *et al.* (2006b,a, 2007a,b); Than *et al.* (2008b)).

1.6 Distinguishing Lineage Sorting from HGT

In Section 1.3 we showed that the gene tree is not always identical to the species tree. Than *et al.* (2007) considered the effect of horizontal gene transfer (HGT) on gene tree under the framework of the coalescent, which we review here. 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 1.5). However, the situation is complicated when

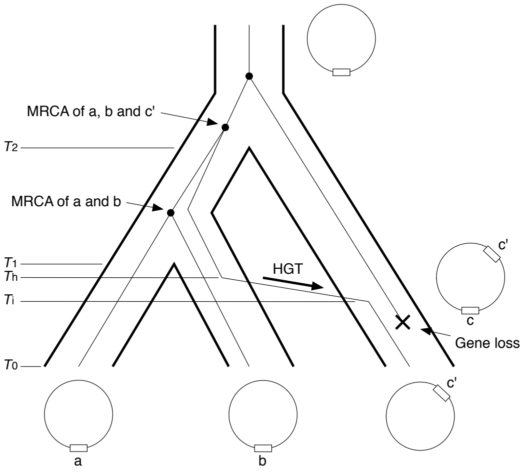

Figure 1.5 *A three species model with a HGT event. A demonstration that a congruent tree could be observed even with HGT.*

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 1.5. 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 1.5. 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 1.5. 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):

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

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

and

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

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 “canceling” such evidence. 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. Recently, Meng & Kubatko (2008) considered a similar scenario involving lineage sorting and hybrid speciation.

1.7 Summary

In the post-genomic era, evidence of massive gene tree incongruence is accumulating in large and diverse groups of organisms. In this chapter, we discussed three major processes that may lead to such incongruence, namely lineage sorting, gene duplication and loss, and reticulate evolution. In the case of the first two processes, the evolutionary history of the set of species still takes the shape of a tree, with gene trees reconciled within the branches of such a tree. However, in the case of reticulate evolution, the evolutionary history of the set of the species’ genomes may be more appropriately modeled by a phylogenetic network. We discussed general approaches for reconciliations under each of the three processes. We also briefly reviewed preliminary work that is being done on extending the coalescent framework in order to enable distinction between lineage sorting as a cause of gene tree incongruence and reticulate evolution as an alternative explanation.

The development of computational tools for identifying gene tree incongruence and inferring species trees despite such incongruence is still in its infancy. There are several major directions for future research in this area, which include, but are not limited to:

1. Developing computational tools for simulating evolution of whole-genomes, or multi-locus data, while incorporating all processes that cause gene tree incongruence. Such tools would play a crucial role in understanding these processes, as well as in enabling the study of the performance of existing and newly developed computational tools.
2. Developing computational tools for reconciling gene trees that can scale to genome-wide data.
3. While methods for addressing lineage sorting already consider multiple loci, computational methods for addressing gene duplication/loss and reticulate evolution mostly work with a pair of trees. It is important that methods are developed for handling multiple trees.
4. We expect that development of sound, unified frameworks for simultaneously

analyzing all three processes and distinguishing among them will be central to progress in this area.

1.8 Acknowledgments

This work was supported in part by DOE grant DE-FG02-06ER25734, NSF grant CCF-0622037, and grant R01LM009494 from the National Library of Medicine. The contents are solely the responsibility of the author and do not necessarily represent the official views of the DOE, NSF, National Library of Medicine or the National Institutes of Health.

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)*.

Beiko, R.G., & Hamilton, N. 2006. Phylogenetic identification of lateral genetic transfer events. *BMC Evolutionary Biology*, **6**.

Bergthorsson, U., Adams, K.L., Thomason, B., & Palmer, J.D. 2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature*, **424**, 197–201.

Bergthorsson, U., Richardson, A., Young, G.J., Goertzen, L., & Palmer, J.D. 2004. Massive horizontal transfer of mitochondrial genes from diverse land plant donors to basal angiosperm Amborella. *Proc. Nat'l Acad. Sci., USA*, **101**, 17747–17752.

Brussow, H., Canchaya, C., & Hardt, W-D. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews*, **68**(3), 560–602.

Burke, J., Schneider, D., & Westpheling, J. 2001. Generalized transduction in *Streptomyces coelicolor*. *Proc. Nat'l Acad. Sci., USA*, **98**, 6289–6294.

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*.

Degnan, J.H., & Rosenberg, N.A. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genetics*, **2**, 762–768.

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*.

Edwards, Scott V., Liu, Liang, & Pearl, Dennis K. 2007. High-resolution species trees without concatenation. *PNAS*, **104**, 5936–5941.

Ellstrand, N.C., Whitkus, R., & Rieseberg, L.H. 1996. Distribution of spontaneous plant hybrids. *Proc. Nat'l Acad. Sci., USA*, **93**(10), 5090–5093.

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.

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*.

Graur, D., & Li, W-H. 2000. *Fundamentals of Molecular Evolution*. Sinauer.

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.

Hodgson, D.A. 2000. Generalized transduction of serotype 1/2 and serotype 4b strains of *Listeria monocytogenes*. *Mol. Microbiol.*, **35**, 312–323.

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.

Jin, G., Nakhleh, L., Snir, S., & Tuller, T. 2006a. Efficient parsimony-based methods for phylogenetic network reconstruction. *Bioinformatics*, **23**, e123–e128. Proceedings of the European Conference on Computational Biology (ECCB 06).

Jin, G., Nakhleh, L., Snir, S., & Tuller, T. 2006b. Maximum likelihood of phylogenetic networks. *Bioinformatics*, **22**(21), 2604–2611.

Jin, G., Nakhleh, L., Snir, S., & Tuller, T. 2007a. Inferring phylogenetic networks by the maximum parsimony criterion: a case study. *Molecular Biology and Evolution*, **24**(1), 324–337.

Jin, G., Nakhleh, L., Snir, S., & Tuller, T. 2007b. A New Linear-time Heuristic Algorithm for Computing the Parsimony Score of Phylogenetic Networks: Theoretical Bounds and Empirical Performance. *Pages 61–72 of:* Mandoiu, I., & Zelikovsky, A. (eds), *Proceedings of the International Symposium on Bioinformatics Research and Applications*. Lecture Notes in Bioinformatics, vol. 4463.

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.

Kubatko, L. S., & Degnan, J. H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.*, **56**, 17–24.

Kuo, Chih-Horng, Wares, John P., & Kissinger, Jessica C. 2008. The Apicomplexan whole-genome phylogeny: An analysis of incongurence among gene trees. *Mol. Biol. Evol.*, **25**(12), 2689–2698.

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.

Linder, C.R., & Rieseberg, L.H. 2004. Reconstructing patterns of reticulate evolution in plants. *American Journal of Botany*, **91**, 1700–1708.

Liu, Liang, & Pearl, Dennis K. 2007. Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology*, **56**(3), 504–514.

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*.

MacLeod, D., Charlebois, R.L., Doolittle, F., & Bapteste, E. 2005. Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. *BMC Evolutionary Biology*, **5**.

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.

Maddison, W.P., & Knowles, L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, **55**(1), 21–30.

Makarenkov, V. 2001. T-REX: Reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, **17**(7), 664–668.

Mallet, J. 2005. Hybridization as an invasion of the genome. *TREE*, **20**(5), 229–237.

Mallet, J. 2007. Hybrid speciation. *Nature*, **446**, 279–283.

Meng, C., & Kubatko, L.S. 2008. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology*. In press.

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.

Mower, J.P., Stefanovic, S., Young, G.J., & Palmer, J.D. 2004. Gene transfer from parasitic to host plants. *Nature*, **432**, 165–166.

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., Jin, G., Zhao, F., & Mellor-Crummey, J. 2005a. Reconstructing phylogenetic networks using maximum parsimony. *Pages 93–102 of: Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference (CSB2005)*.

Nakhleh, L., Ruths, D., & Wang, L.S. 2005b. 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).

Noor, M.A.F., & Feder, J.L. 2006. Speciation genetics: evolving approaches. *Nature Review Genetics*, **7**, 851–861.

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

Ohno, S. 1970. *Evolution by gene duplication*. Springer.

Olson, M.V. 1999. When less is more: gene loss as an engine of evolutionary change. *American Journal of Human Genetics*, **64**, 18–23.

Page, R. D. M., & Charleston, M. A. 1997a. 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. 1998. GeneTree: Comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, **14**(9), 819–820.

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

Pollard, D. A., Iyer, V. N., Moses, A. M., & Eisen, M. B. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet.*, **2**, 1634–1647.

Posada, D., & Crandall, K.A. 2002. The effect of recombination on the accuracy of phylogeny estimation. *J. Mol. Evol.*, **54**(3), 396–402.

Posada, D., Crandall, K.A., & Holmes, E.C. 2002. Recombination in Evolutionary Genomics. *Annu. Rev. Genet.*, **36**, 75–97.

Rieseberg, L.H., & Carney, S.E. 1998. Plant hybridization. *New Phytologist*, **140**(4), 599–624.

Rieseberg, L.H., Baird, S.J.E., & Gardner, K.A. 2000. Hybridization, introgression, and linkage evolution. *Plant Molecular Biology*, **42**(1), 205–224.

Rokas, A., Williams, B.L., King, N., & Carroll, S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 798–804.

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).

Schicklmaier, P., & Schmieger, H. 1995. Frequency of generalized transducing phages in natural isolates of the *Salmonella typhimurium* complex. *Appl. Environ. Microbiol.*, **61**, 1637–1640.

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).*

Stewart, G.J., & Sinigalliano, C.D. 1991. Exchange of chromosomal markers by natural transformation between the soil isolate, Pseudomonas JM300 and the marine isolate, Pseudomonas stutzeri strain ZoBell. *Antonie Van Leeuwenhoek*, **59**(1), 19–25.

Syring, J., Willyard, A., Cronn, R., & Liston, A. 2005. Evolutionary relationships among *Pinus* (Pinaceae) subsections inferred from multiple low-copy nuclear loci. *Am. J. Bot.*, **92**, 2086–2100.

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.

Than, C., & Nakhleh, L. 2008. SPR-based tree reconciliation: Non-binary trees and multiple solutions. *Pages 251–260 of: Proceedings of the Sixth Asia Pacific Bioin-*

formatics Conference (APBC).

Than, C., & Nakhleh, L. 2009. *Efficient Genome-scale Inference of Species Trees by Minimizing Deep Coalescences*. Under review.

Than, C., Ruths, D., Innan, H., & Nakhleh, L. 2007. Confounding factors in HGT detection: Statistical error, coalescent effects, and multiple solutions. *Journal of Computational Biology*, **14**(4), 517–535.

Than, C., Sugino, R., Innan, H., & Nakhleh, L. 2008a. Efficient Inference of Bacterial Strain Trees From Genome-scale Multi-locus Data. *Bioinformatics*, **24**, i123–i131. Proceedings of the 16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB '08).

Than, C., Jin, G., & Nakhleh, L. 2008b. Integrating Sequence and Topology for Efficient and Accurate Detection of Horizontal Gene Transfer. *Pages 113–127 of: Proceedings of the Sixth RECOMB Comparative Genomics Satellite Workshop*. Lecture Notes in Bioinformatics, vol. 5267.

Than, C., Ruths, D., & Nakhleh, L. 2008c. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, **9**(1), 322.

Welch, R.A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., & *et al.*, J. Hackett. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 17020–17024.

Woegerbauer, M., Jenni, B., Thalhammer, F., Graninger, W., & Burgmann, H. 2002. Natural genetic transformation of clinical isolates of *Escherichia coli* in urine and water. *Appl. Environ. Microbiol.*, **68**(1), 440–443.

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.