Network (Reticulate) Evolution: Biology, Models, and Algorithms

C. Randal Linder
School of Biological Sciences
The University of Texas at Austin
Austin, Texas 78712

Bernard M.E. Moret
Department of Computer Science
The University of New Mexico
Albuquerque, New Mexico 87131

Luay Nakhleh Tandy Warnow
Department of Computer Sciences
The University of Texas at Austin
Austin, Texas 78712

November 5, 2003

1 Introduction

1.1 Phylogeny and Its Centrality to Biology

Phylogenies are the main tool for representing evolutionary relationships among biological entities at the level of species and above. Because in nearly all cases of biological evolution it is impossible to witness the history of speciation events, biologists, mathematicians, statisticians, and computer scientists have designed a variety of methods for the reconstruction of these events, with the usual model being phylogenetic trees. Over the last 30 years, biologists have come to embrace reconstruction of phylogenetic trees as a major research goal [75, 101, 111]—in fact, a casual search of the MEDLINE database for the word "phylogeny" returns approximately 24,000 articles since 1966, with an exponentially growing curve. Currently, one of the major endeavors in biology is reconstructing the complete "Tree of Life"—the phylogeny of all living organisms on Earth. As part of this endeavor, the National Science Foundation set up a research program called "Assembling the Tree of Life," under which it has already made nearly 20 awards totaling over $30 million, mostly to explore in depth one or another particular branch of the tree. In addition, the NSF recently awarded our group a $12 million grant for the development of methodologies and infrastructure for assembling the Tree of Life (see `http://www.phylo.org`). Once completed, this phylogeny will stand as one of science's great accomplishments. Even in its current partially completed form, its implication—that all living things on Earth today (from bacteria, to seaweeds, to mushrooms, to humans) are related—has forever changed our perception of the world around us. The use of phylogenetic principles is almost as

ubiquitous today as the idea of Darwinian evolution. There is growing appreciation that phylogeny is an indispensable interpretive framework for studying evolutionary processes, and indeed is central to the organization and interpretation of information on all characteristics of organisms, from structure and physiology to genomics. With modern data-gathering methods and improved analytical tools, resolution of this dominant aspect of life's history is finally an attainable goal.

Phylogeny, because it reflects the history of transmission of life's genetic information, has unique power to organize our knowledge of diverse organisms, genomes, and molecules. A reconstructed phylogeny helps guide our interpretation of the evolution of organismal characteristics, providing hypotheses about the lineages in which traits arose and under what circumstances, thus playing a vital role in studies of adaptation and evolutionary constraints [74, 144, 149, 158, 167]. Patterns of divergence of species lineages indicated by the tree inform the dynamics of speciation and, to some extent, extinction—the two forces that generate and reduce biodiversity [35, 80]. Phylogeny informs far more than evolutionary biology, however. The evolutionary histories of genes bear the marks of the functional demands to which they have been subjected, so that phylogenetic analyses can elucidate functional relationships within living cells [81, 90, 286]. Pharmaceutical companies are thus increasingly using phylogenetic analyses to make functional predictions from sequence databases of gene families [15], to predict ligands [36], and to help in the development of vaccines [93] and antimicrobials and herbicides [26, 187].

Because phylogenies are such an important part of biological investigations, many methods exist for reconstructing phylogenies. Many of these methods work on aligned biomolecular sequences (i.e., DNA, RNA, or amino-acid sequences), and use these sequences to infer the evolutionary history of the sequences. For the most part, these methods assume that the phylogeny underlying the data is a tree. However, such is not always the case: for many organisms, a significant level of genetic exchange occurs between lineages, and for some groups, lineages can combine to produce new independent lineages. These exchanges and combinations transform a tree into a network.

1.2 The "Tree of Life" Is Not Really a Tree

Ford Doolittle [59] famously wrote

> Molecular phylogeneticists will have failed to find the "true tree," not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot properly be represented as a tree.

Indeed, events such as meiotic and sexual recombination, horizontal gene transfer and hybrid speciation cannot be modeled by bifurcating trees. Meiotic recombination occurs in every generation at the level of individual chromosomes; sexual recombination commonly acts at the population level and recombines the evolutionary histories of genomes. Hybrid speciation is very common in some very large groups of organisms: plants, fish, frogs, and many lineages of invertebrates, and horizontal gene transfer is ubiquitous in bacteria. Although the mixing (reticulation) of evolutionary histories has long been widely appreciated

and acknowledged, there has been comparatively little work on computational methods for studying and estimating reticulate evolution, especially at the species level.

In this survey, we address the biological aspects of reticulate evolution, present existing mathematical models, and discuss current computational approaches. We begin with an overview of phylogenetic tree reconstruction before addressing the specifics of reticulate evolution.

2 Phylogenetic Tree Reconstruction

2.1 Data for Phylogeny Reconstruction

Phylogenies are most commonly reconstructed using aligned biomolecular sequences (DNA, RNA, or amino acid) for particular genes or non-coding regions of DNA—although whole-genome data (gene content and gene order on chromosomes) are getting more common and offer complementary attributes. The evolutionary history of the species from which the sequences were obtained is then inferred by examining how individual sites (positions) within the sequences evolve on each candidate tree; the tree(s) that show the most reasonable evolutionary scenario under a particular model of evolution are returned. What is reconstructed is the evolutionary history of a gene tree (or more precisely the evolutionary history of a particular region of DNA), which need not coincide with the history of the species [150].

Usually, researchers want to reconstruct the phylogeny of groups of organisms that share a most recent common ancestor (MRCA), termed a *monophyletic group* or a *clade*. DNA sequences that have evolved from a single MRCA at the root of a clade are said to be *orthologous*, in contrast to DNA sequences that are the result of gene duplications or alleles that evolved prior to the MRCA of the clade, which are said to be *paralogous*. In the case of DNA sequences, gene trees that are reconstructed from orthologs will be identical to the species trees, but it is not always possible to be certain that all of the gene sequences used for phylogenetic reconstruction are orthologous. When paralogs are mistakenly used for reconstructing the gene tree, the species tree inferred from the gene tree will usually be incorrect. If all of the orthologs are present in the extant taxa from which the DNA sequences are taken, then the use of paralogs in tree reconstruction can be ameliorated by more extensive sampling of the species. However, a number of population genetic processes can cause orthologs to be randomly or systematically lost in some species (*lineage sorting*): genetic drift and population bottlenecks (random), and natural selection (systematic). Thus, when a species lacks a particular ortholog, it is possible to use a paralog and be unable to detect the mistake. For those cases where orthologs are lost at random, use of multiple DNA regions can minimize the problem by revealing patterns of relationships reconstructed at rates greater than expected by chance. To avoid problems generated by natural selection, sequences from non-coding regions can be sought, since these usually evolve more randomly than coding sequences, which are under selection. However, this option is not always available because non-genic DNA sequences usually evolve too rapidly to be aligned reliably for evolutionarily distant species. Determining whether DNA sequences are orthologous in distantly related species is a current topic of research.

Recovering the correct tree also depends on using a DNA sequence that evolves at an appropriate rate. A sequence that evolves too rapidly relative to the times of divergence for a clade may be unalignable. Alternatively a sequence that evolves too slowly will lack sufficient information to infer the species relationships with confidence. When the sequences from a gene have insufficient variation to resolve phylogenetic relationships, sequences from two or more genes are sometimes concatenated into a *combined analysis*; a phylogeny is then sought for the combined dataset. Since not all datasets necessarily evolve under the same tree (as noted earlier), combined analyses should only be attempted when the researcher has evidence that the different sequences in the combined analyses have the same evolutionary history (i.e., the same tree) or when sufficiently many are combined that any incompatible gene history disappears under the mass of compatible information. Determining when it is safe to combine datasets is an important part of a phylogenetic analysis, and several techniques have been proposed to address this problem [31, 40, 47, 53, 109, 185, 277]. However, incompatibilities between trees can arise for a number of reasons in addition to lineage sorting. One of these is reticulate evolution - see [230] for a discussion of these issues.

2.2 Evaluating Reconstruction Methods

There are many phylogenetic reconstruction methods, the majority of which are attempts to solve NP-hard optimization problems. Because accuracy in evolutionary reconstruction cannot usually be determined for real datasets, the accuracy of a phylogeny reconstruction method is studied under the assumption that the input to the reconstruction method is a set of sequences that is generated by an unknown but fixed *model tree* from some given Markov model of sequence evolution (see Section 2.3). This assumption allows methods to be explored with respect to the conditions under which they will return accurate estimations of the tree, either through mathematical theorems, or through simulation. These studies, both mathematical and simulation-based, have been very influential in systematic biology; for example, mathematical studies have established that methods such as maximum parsimony and maximum compatibility, even when given arbitrarily long sequences, do not always reconstruct accurate phylogenies, whereas other methods, such as neighbor joining and maximum likelihood, will do so under most models, if given sufficiently long sequences [76]. Simulation studies have been particularly helpful in exploring performance under finite sequence lengths, and have become the main tool for the evaluation of reconstruction methods [102, 103].

2.3 Stochastic Models of DNA Sequence Evolution

Both approaches (mathematics and simulation) to evaluating performance require that the stochastic model of evolution be specified. Many stochastic models of DNA sequence evolution have been proposed; most assume that the sites (positions within the sequences) evolve down a tree via nucleotide substitutions under a Markov process. The stochastic process under which an individual site evolves down an edge within the tree can therefore be given by a 4×4 substitution matrix M_e for that edge e, dictating the probability of each

of the possible states at the "bottom" of the edge as a function of the state at the "top" of the edge. Thus, the evolution of a single site down the tree is given by a rooted tree with edges annotated with substitution matrices.

The simplest of these models is the Jukes-Cantor [122] model, which assumes:

1. the sites evolve identically and independently (the *iid* assumption),
2. the state of each site at the root is randomly selected, and
3. if a site changes state on an edge, it changes with equal probability to each of the remaining three states.

The Jukes-Cantor model thus has only one free parameter on each edge.

Other models of site substitution with more free parameters are considered to be biologically more realistic. The simplest of these is the Kimura 2-parameter (K2P) model [127]; the most general is the General Markov model, which allows the substitution matrices to be quite general. Many (but not all) phylogenetic reconstruction methods are guaranteed to be *statistically consistent* under the General Markov model (and hence under all the models it contains)—that is, these methods can infer the true tree with high probability when given long enough sequences. Statistical consistency is still possible under models where these conditions are relaxed in a controlled fashion—for example, by allowing the sites to have rates of evolution that vary with the sites and that are drawn from a known distribution. However, if the rates are drawn from an unknown distribution (for example, if an unknown proportion of sites are constant), then it may not be possible to identify the true tree, even from infinite data. See [76, 126, 143] for a further discussion of these issues.

The *iid* models of DNA site substitution are clearly unreasonable, yet most enhance models simply allow rates to vary across sites, which by itself does not alter the iid assumption. Bruno and Halpern [27] have developed a model for protein-coding DNA sequences that addresses most of the weaknesses. In particular, they use the GM model to generate changes, but then have these changes fixed or lost according to the equilibrium frequencies of the amino acid that would result from the changes. The model is extensible to other types of sequences.

Similar, but necessarily more complex, models can be built from larger sequence units: from codons, for instance, one can build a mode of amino-acid evolution, using, among other things, a 20×20 substitution matrix.

2.4 Simulation Studies

Phylogenetic reconstruction methods are evaluated according to two basic types of criteria: *statistical performance*, which addresses the accuracy of the method under a specified stochastic model of evolution, and *computational performance*, which addresses the computational requirements of the method. A method is said to be *statistically consistent* with respect to a specific model of evolution if it is guaranteed to recover the true tree with probability going to 1 as the amount of data (i.e., sequence length) goes to infinity.

Accuracy in a phylogenetic reconstruction method is determined primarily by comparing the unrooted leaf-labelled tree obtained by the method to the "true" tree. Since the true

tree is usually unknown, accuracy is addressed either theoretically, with reference to a fixed but unknown tree in some model of evolution, or through simulation studies.

In simulation studies, a model phylogeny is generated, then a set of sequences is evolved down the edges of the the model phylogeny according to some chosen model of sequence evolution, and the sequences thus obtained at the leaves are given as input to the reconstruction method under study. The resulting phylogeny is compared to the model tree to assess the topological accuracy of the reconstruction. Of the several measures proposed for quantifying the topological accuracy of tree reconstruction methods, the most commonly used is the Robinson-Foulds (RF) measure [227], which we now describe.

The Robinson-Foulds measure Every edge e in a leaf-labeled tree T defines a bipartition π_e on the leaves (induced by the deletion of e), so that we can define the set $C(T) = \{\pi_e : e \in E(T)\}$, where $E(T)$ is the set of all internal edges of T; this is called the *character encoding* of the tree T. If T is a model tree and T' is the tree inferred by a phylogenetic reconstruction method, we define the *false positives* to be the edges of the set $C(T') - C(T)$ and the *false negatives* to be those of the set $C(T) - C(T')$. We can then compute the error *rates* by normalizing these values by the number of internal edges in the model tree; since we assume that model trees are binary, this is $n-3$ for n-leaf trees. We obtain:

- The *false positive rate (FP)* is $\dfrac{|C(T') - C(T)|}{n-3}$.
- The *false negative rate (FN)* is $\dfrac{|C(T) - C(T')|}{n-3}$.

When both trees are binary, we have $FP = FN$; in general, we have $FP \leq FN$. Since n is the number of internal edges of a rooted binary tree on n leaves, the false positive and false negative rates are values in the range $[0,1]$. The RF distance between T and T' is simply the average of these two rates, i.e., $\frac{FN+FP}{2}$. (An equivalent formulation is $RF(T,T') = \frac{1}{2}|C(T) \triangle C(T')|$, where $\triangle$ denotes the symmetric difference.) Error rates below 10% are considered not too bad, but systematists prefer to see error rates below 5%.

Simulation studies show that the topological accuracy of a method is affected by the number of leaves in the tree, the maximum evolutionary diameter, the deviation from a molecular clock, as well as tree shape [102, 103, 176, 177, 175, 172], and that different methods are affected differently by these aspects of the model tree. Other simulation studies show that different heuristics for maximum parsimony or maximum likelihood also perform differently [278, 232].

2.5 Phylogenetic Tree Reconstruction Methods

There are three basic types of phylogenetic reconstruction methods in common use: distance-based methods, maximum parsimony heuristics, and maximum likelihood heuristics. Generally, only the distance-based methods operate in polynomial time, since the other methods attempt to solve NP-hard optimization problems. (Distance-based methods may also

attempt to solve NP-hard optimization problems, but there are polynomial-time distance-based methods, such as *neighbor joining (NJ)* [233], that perform very well in practice and that do not attempt to solve NP-hard optimization problems.)

Distance-based methods Distance-based methods operate by first estimating pairwise distances and then computing an edge-weighted tree using those distances. Such methods are guaranteed to reconstruct the true tree if their estimates of pairwise distances are sufficiently close to the number of evolutionary events between pairs of taxa [126]. For many models of biomolecular sequence evolution, estimation of sufficiently accurate pairwise distances is possible [143]; for example, log-det [255] distances are statistically consistent estimators for the General Markov model of evolution. Therefore, distance-based methods are statistically consistent for the General Markov model of DNA sequence evolution, and hence for its constituent submodels. Distance-based methods are typically very fast: most run in $O(n^3)$ time, where n is the number of taxa. Of the various distance-based methods in use, neighbor joining [233] is certainly the most popular.

Maximum parsimony heuristics This approach is typical in the analysis of DNA sequences, in which the objective is the tree with the minimum number of nucleotide substitutions across the tree.

- *Input:* Set S of aligned DNA sequences
- *Output:* Tree T leaf-labelled by S and with internal nodes labelled by additional DNA sequences, so that $\sum_e \text{Hamming}(e)$ is minimized, where $\text{Hamming}(e)$ is the number of sites differing between the sequences labelling the endpoints of the edge e and the sum ranges over the edges e in the tree T.

The maximum parsimony problem is NP-hard [78]. Due to the popularity of the maximum parsimony criterion in phylogeny reconstruction, many heuristics have been devised to reconstruct locally optimal trees (see [76] for a review of these heuristics). These heuristics operate by hill climbing through an exponentially-sized tree space. Maximum parsimony is not a statistically consistent technique for estimating phylogenies, even under the simplest 4-state (Jukes-Cantor) model [73]; nevertheless, it is a very popular approach in systematics because it is more computationally efficient with large numbers of taxa than maximum likelihood.

Maximum likelihood Another major criterion for phylogeny reconstruction from molecular sequences is *Maximum Likelihood* (ML), which was pioneered by Felsenstein, Edwards, and Thompson. In the ML problem, we seek the tree T and its associated parameters (such as edge-lengths, rates of evolution for each site, etc.) that maximize the probability of generating the given set of sequences. The general idea behind maximum likelihood estimators is the observation that

$$P(Model|Data) = \frac{P(Model\ and\ Data)}{P(Data)} = \frac{P(Data|Model)P(Model)}{P(Data)}.$$

In this formulation, $P(Model|Data)$ is proportional to $P(Data|Model)$; therefore, we can justify estimating a model by finding the model that maximizes the conditional probability $P(Data|Model)$, which is also called the *likelihood* of the data.

From a practical standpoint, the ML problem is difficult on two levels: first finding the best leaf-labelled tree and then finding the edge parameters for that tree. In practice, hill climbing heuristics are used for both optimization problems (setting edge parameters, and finding the best tree), and heuristics for ML are generally slower than heuristics for MP. If solved exactly, however, ML is statistically consistent under the General Markov model of evolution, and therefore under all of its submodels. Bayesian approaches are also used; these seeks to estimate parameters in much the same way, but the goal is the recovery of the conditional posterior probabilities, using some initial estimate (or guess) for the *a priori* probability.

Genomic rearrangements Genomic changes in gene content and gene order are rare events [229]. Examples include intron gains/losses [273], retroposon integrations [245], signature sequences [54, 84], genetic code variants [124, 267], changes in gene order [206], and gene duplications [159]. Changes in gene order and gene duplications are examples of a more general class of events, *genomic rearrangements*. Plant biologists have used structural changes in chloroplast genomes to infer phylogenies for nearly 20 years [63, 206, 121, 193, 206, 257, 258], but only with the advent of whole-genome sequencing can they be used on a large scale. The maximum parsimony problem can be extended to any kind of data that evolves via a precise set of mutations: given the set S of taxa, find the tree in which the minimum number of evolutionary events (or weighted sum thereof) occurs. Thus parsimony can be used for whole-genome evolution. Synteny (colocation on the same chromosome) was famously used by Nadeau and Taylor in predicting the number of genes in man and mouse from rudimentary data [174]. Synteny and gene order are some of the best data for use in orthology analysis (see, e.g., [12, 50, 171]).

3 Biology of Reticulate Evolution

Reticulation in biology refers to the lack of independence between two evolutionary lineages. In effect, when reticulation occurs, two or more independent evolutionary lineages are combined at some level of biological organization. Because life is organized hierarchically, reticulation can occur at different levels: chromosomes, genomes and species.

At the species level, events such as hybrid speciation (by which two lineages recombine to create a new one) and horizontal gene transfer (by which genes are transferred across species) are the main causes of reticulate evolution. Within each lineage, at the population level, sexual recombination causes evolution to be reticulate, whereas meiotic recombination causes the shuffling of genes at the chromosomal level. Figure 1 illustrates these three scenarios. The phylogeny in Figure 1(a) depicts a hybrid speciation scenario, in which species C is the product of hybridization between B and D. Zooming in on a lineage of the phylogeny gives a picture of reticulate evolution at the population level, as in Figure 1(b). Finally, zooming in on an individual in each population, reticulation between

chromosomes can be viewed, as in Figure 1(c). Looking through a macroevolution lens

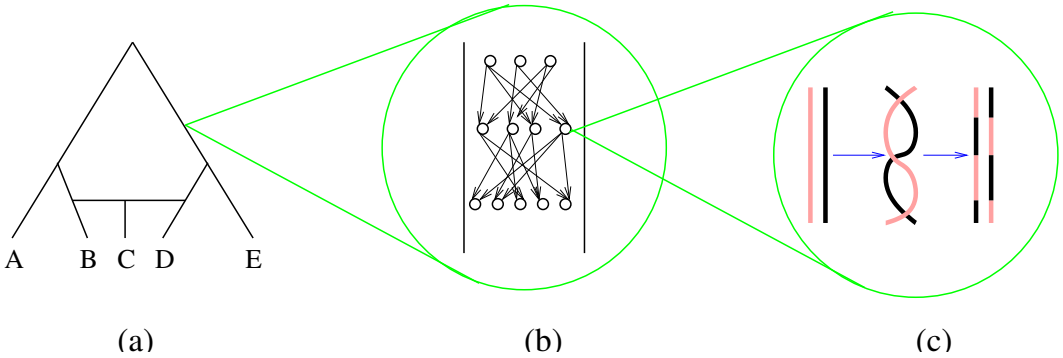

Figure 1: Reticulation at various levels: (a) species level, (b) population level, and (c) individual (chromosomal) level.

(evolution among lineages), only reticulate events at the species level fail to be modeled by a bifurcating tree. However, looking through a microevolution lens (evolution within a lineage), sexual and meiotic recombination fail to be modeled by a bifurcating tree. Since phylogenies are usually constructed at either the population or the species level, meiotic recombination does not cause a species-level reticulate evolutionary history, but it can confound species-level inference of reticulation by producing patterns that have the appearance of species-level reticulation. We now briefly explain reticulation at the chromosomal, population and species levels.

Reticulation between chromosome pairs: meiotic recombination During each round of sexual reproduction, the total number of chromosomes must be halved to produce the gametes. The process is called meiosis and during one phase of it the chromosome pairs (sister chromatids) exchange pieces in a precise fashion known as meiotic recombination. The net result is chromatids that have two or more evolutionary histories on them. Blocks of chromosomes that share a single evolutionary history are referred to as haplotype blocks.

Reticulation within a lineage: sexual recombination For sexually reproducing organisms, there is recombination of nuclear genomes during each bout of reproduction. Each parent contributes half of its original nuclear genome—one sister chromatid from each chromosome—and each of these chromosomes have themselves undergone meiotic recombination during the process of producing the haploid gametes (sex cells). Because different parts of each parent's contribution to the genome of the next generation may have a different evolutionary history from that of the other parent's contribution, sexual recombination is a form of population-level reticulation. Organellar genomes (mitochondria and chloroplasts) are usually inherited uniparentally so they do not usually undergo any sort of sexual recombination.

Reticulation among lineages: horizontal gene transfer and hybrid speciation In horizontal (also called lateral) gene transfer (HGT for short), genetic material is transferred

from one lineage to another; see Figure 2(a). In an evolutionary scenario involving horizontal transfer, certain sites (specified by a specific substring within the DNA sequence of the species into which the horizontally transferred DNA was inserted) are inherited through horizontal transfer from another species (as in Figure 2(b)), while all others are inherited from the parent (as in Figure 2(c)). Thus, *each site evolves down one of the trees contained inside the network*.

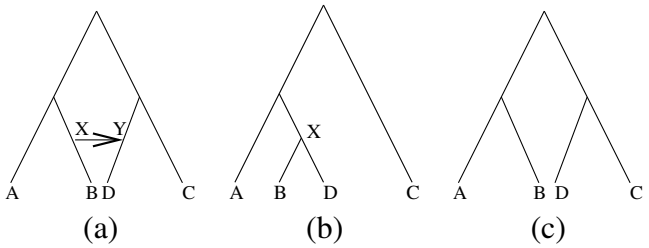

Figure 2: Horizontal gene transfer: the species network in (a) and the two possible gene trees in (b) and (c).

Horizontal gene transfer is of paramount importance in the study of the evolution of prokaryotes and, to a lesser extent, the eukaryotic lineages that have interspecific hybridization. Describing the ubiquity of HGT, de la Cruz and Davies [52] wrote

> It is clear that genes have flowed through the biosphere, as in a global organism. Horizontal gene transfer, once solely of interest for practical applications in classical genetics and biotechnology, has now become the substance of evolution.

Horizontal transfers are believed to be ubiquitous among bacteria and still quite common in other branches of the "tree"—although this view has recently been challenged [67, 234, 235, 134]. Three mechanisms of HGT in the Archaea and Bacteria are *transformation* (uptake of naked DNA from the environment), *conjugation* (transfer of DNA by direct physical interaction between a donor and a recipient), and *transduction* (transfer of DNA by phage infection) [197].

The second non-treelike event acting at the species level is hybrid speciation. In hybrid speciation, two lineages recombine to create a new species; see Figure 3(a) for a visual rep-

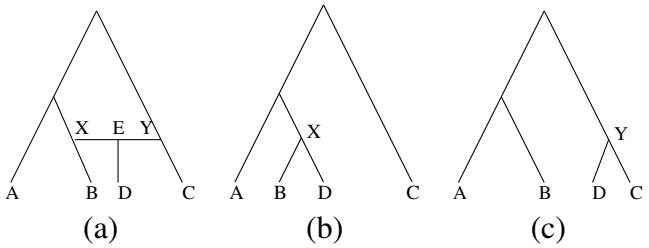

Figure 3: Hybrid speciation: the species network in (a) and the two possible gene trees in (b) and (c).

resentation of this scenario. The new species may have the same number of chromosomes as its parent (*diploid hybridization*) or the sum of the number of its parents (*polyploid hybridization*).

Hybrid speciation comes about in at least three ways: allopolyploidization, autopolyploidization, and diploid or homoploid hybrid speciation. Autopolyploidization is probably more properly considered a specialized form of normal (bifurcating) speciation since only a single parental species is involved in its production. Allopolyploidization is hybrid speciation between two species resulting in an offspring that has the complete diploid chromosome complement of both its parents. Each parent need not have the same number of chromosomes. Allopolyploidization results in instantaneous speciation because any backcrossing to the diploid parents would produce inviable or sterile triploid offspring. Diploid hybrid speciation is a normal sexual event where each gamete has a haploid complement of the chromosomes from its parent, but the gametes that form the zygote come from different species. In nearly all cases, both parents must have the same number of chromosomes. In this case, successful backcrossing to the parents is possible, so it is thought that the hybrids have to be isolated from the parents by undergoing selection for life in a novel environment [214]. Not surprisingly, diploid hybrid speciation is much rarer than polyploidization.

Consider how an individual site evolves down a network modelling hybrid speciation. For normal diploid organisms, each chromosome consists of a pair of homologs. In a diploid hybridization event, the hybrid inherits one of the two homologs for each chromosome from each of its two parents. Since homologs assort at random into the gametes (sex cells), each has an equal probability of ending up in the hybrid. In polyploid hybridization, both homologs from both parents are contributed to the hybrid. Prior to the hybridization event, each site on the homolog has evolved in a tree-like fashion, although due to meiotic recombination (exchanges between the parental homologs during production of the gametes), 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 contained inside the network representing the hybridization event (see Figures 3(b) and 3(c)). As in the case of HGT, *each site evolves down one of the trees contained inside the network*.

Hybrid speciation is very common in some groups of organisms (plants, fish, amphibians, some groups of invertebrates, and possibly fungi) and is virtually absent in others, notably mammals and most arthropods. These latter groups do produce very occasional hybrids, but they are usually triploid and are only able to survive by asexual reproduction. Odd ploidy levels cannot reproduce sexually because the odd number of chromosome sets does not allow correct pairing during meiosis. The resulting gametes have unequal numbers of chromosomes and are nearly always inviable. These asexual lineages are considered evolutionary dead ends and are expected to be short lived. The reasons some groups can successfully speciate via hybridization while others cannot is not understood (see [186] for a review). It is thought that developmental complexity may play an important role: animals with complex developmental programs may simply be unable to produce viable offspring after interspecific mating has taken place. It has also been argued that species with chromosomal sex determination (species with distinct sex chromosomes) are much less likely to produce hybrid species. However, the true reasons for the barrier(s) to successful hybrid

speciation have yet to be demonstrated.

4 Mathematical Models of Reticulate Evolution

4.1 Population Level

Strimmer *et al.* [260] proposed DAGs (directed acyclic graphs) as a model for describing the evolutionary history of a set of sequences under recombination events. They also described a set of properties that a DAG must possess in order to provide a realistic model of recombination. Strimmer *et al.* [261] proposed adopting *ancestral recombination graphs* (ARGs), due to Hudson [108] and Griffiths and Marjoram [89] as a more appropriate model of phylogenetic networks. ARGs are rooted graphs that provide a way to represent linked collections of clock-like trees by a single network. Another network-like model is pedigrees, designed to represent the parentage of individual organisms—so that the indegree of each internal node of a pedigree is either 0 or 2, thereby not allowing for tree nodes; Figure 4 gives an example of a pedigree, where squares represent males, and circles represent females. Nodes of indegree 2 correspond to recombination events.

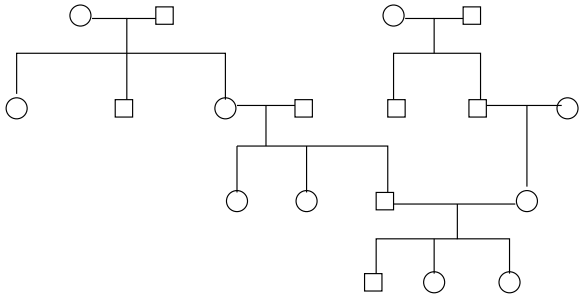

Figure 4: An example of a pedigree.

4.2 Species Level

Hallett and Lagergren [95] described a set of conditions on rooted DAGs to use them as models for evolution under lateral transfer events. In [145], Linder *et al.* proposed a model of phylogenetic networks that is also based on DAGs. Like Strimmer *et al.* [260], Linder *et al.* use DAGs to describe the topology of phylogenetic networks and, like Hallett and Lagergren [95], they add a set of (mostly simpler) conditions to ensure that the resulting DAGs reflect the properties of reticulation. We now review the mathematical model of phylogenetic networks [145].

4.2.1 Model networks

A phylogenetic network $N = (V,E)$ is a rooted DAG obeying the following constraints. The set V of nodes is partitioned into two sets:

- V_T: the set of **tree nodes**. A node $v \in V$ is a tree node if and only if one of these three conditions holds:
  - $indegree(v) = 0$ and $outdegree(v) = 2$: v is the root,
  - $indegree(v) = 1$ and $outdegree(v) = 0$: v is a leaf, or
  - $indegree(v) = 1$ and $outdegree(v) = 2$: v is an internal node.
- V_N: the set of **network nodes**. A node $v \in V$ is a network node if and only if one of these two conditions holds:
  - $indegree(v) = 2$ and $outdegree(v) = 1$: the species at node v is the product of homoploid or allo-polyploid hybridization, or
  - $indegree(v) = 1$ and $outdegree(v) = 1$: the species at node v is the product of auto-polyploidization.

We clearly have $V_T \cap V_N = \emptyset$ and can easily verify that we have $V_T \cup V_N = V$.

The set E of edges is partitioned into two sets:

- E_T: the set of **tree edges**. An edge $e = (u,v) \in E$ is a tree edge if and only if v is a tree node.
- E_N: the set of **network edges**. An edge $e = (u,v) \in E$ is a network node 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. For any pair of nodes u and v in V, if (u,v) is an edge in E, then at least one of u or v is a tree node. Figure 5 shows an example of a phylogenetic network in which the species at node Z is the product of (homoploid or allo-polyploid) hybridization and the species at node W is the product of auto-polyploidization. Dashed edges denote network edges and solid edges denote tree edges.

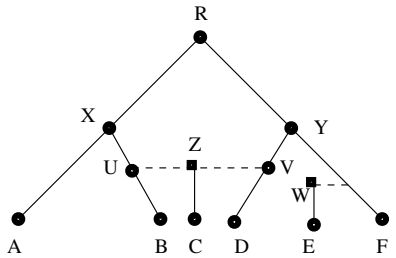

Figure 5: A phylogenetic network N on 6 species. The solid circles denote the tree nodes, and the solid squares denote the network nodes. The solid lines denote the tree edges, and the dashed lines denote the network edges.

In such a network, a species appears as a directed path p that does not contain any network edge. If p_1 and p_2 are two directed paths that define two distinct species, then p_1 and p_2 must be edge-disjoint, that is, the two paths cannot share edges. For example, the

directed path p from node X to node B in Figure 5 could define species B (as could the path from R to B), whereas the directed path from node X to node C does not define a species, since it contains a network edge.

A phylogenetic network $N = (V,E)$ defines a partial order on the set V of nodes. Based on this partial order, we assign times to the nodes of N, associating time $t(u)$ with node u. If there is a directed path p from node u to node v, such that p contains at least one tree edge, then we must have $t(u) < t(v)$ (in order to respect the time flow). If $e = (u,v)$ is a network edge, then we must have $t(u) = t(v)$ (because hybridization is, at the scale of evolution, an instantaneous process). Because of that property, the orientation of a network edge is irrelevant when examining time flows.

Given a network N, we say that p is a *positive-time directed path* from u to v, if p is a directed path from u to v, and p contains at least one tree edge. Given a network N, two nodes u and v cannot co-exist in time if the following condition holds:

- 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}.

Notice that the simple condition of having a positive-time directed path from u to v is, by itself, not sufficient to capture all temporal constraints imposed by reticulation events; Figure 6 illustrates this point. In Figure 6, $t(Y) = t1$ and $t(X) = t4$; further, hybridization events $H1$ and $H2$ occur at times $t2$ and $t3$, respectively. It is obvious that the two hybridization events imply $t1 < t2 < t3 < t4$, which, in turn, implies that X and Y cannot co-exist in time, and hence cannot be the "parents" of a hybridization event. Obviously, there does not exist a positive-time directed path from Y to X. Yet, there exists a sequence $P = \langle p_1, p_2, p_3 \rangle$ of positive-time directed paths, where p_1 is the directed path from Y to A, p_2 is the directed path from B to C, and p_3 is the directed path from D to X; further, $H1$ is a

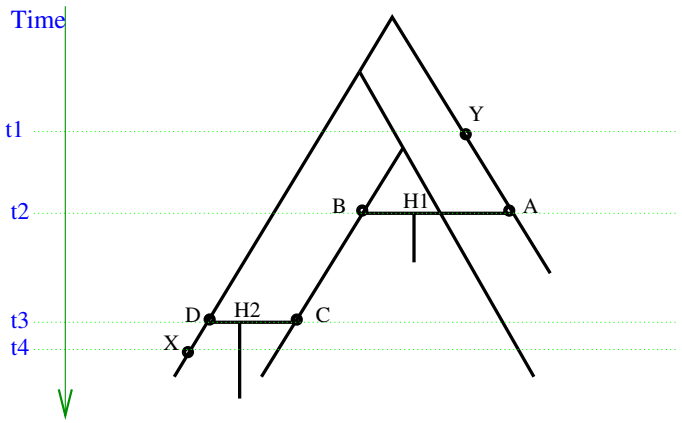

Figure 6: A scenario illustrating two nodes X and Y that cannot co-exist in time.

network node whose two parents are A and B, and $H2$ is a network node whose two parents are C and D. Hence, according to our definition, X and Y cannot co-exist in time. Since events such as hybridization and lateral gene transfer occur between two lineages (nodes in the network) that co-exist in time [150, 192], a phylogenetic network N must satisfy the following property:

- If two nodes x and y cannot co-exist in time, then they cannot hybridize, that is, there cannot exist node v with (x,v) and (y,v) in E.

5 Detection and Reconstruction of Reticulate Evolution: What Needs to be Done

As we saw in phylogenetic tree inference, the design and analysis of methods for the reconstruction of phylogenetic trees has several components:

- software for simulation studies: generating random model trees, simulating evolution down trees, and comparing inferred trees to model trees for accuracy
- algorithms and software for reconstructing phylogenetic trees

The phylogenetics community has produced much along these lines, and many of the best tools are quite good - provided the datasets are not too large. Similarly, we also need software for simulation studies for reticulate evolution, and algorithms and software for detecting reticulation and for reconstructing networks. Unfortunately, despite some progress along these lines, still very much needs to be done. In this section we describe the current state of these research agenda.

5.1 Detecting Reticulate Evolution

There are several methods which attempt to detect reticulation, and a few methods for reconstructing reasonable reticulate scenarios when reticulate evolution is suspected; however, much research needs to be done in order to develop better methods for both problems. The main challenge in detecting reticulation is that patterns in the data that are suggestive of reticulation may also be present due to other factors (such as lineage sorting, inadequate data, insufficient analyses, etc.), and it is difficult to distinguish between these different conditions; see [230] for a discussion of these issues. For this reason, among others, current methods for determining that reticulation has occurred have not been entirely successful. Reconstruction methods are also generally not yet sufficiently accurate for phylogeny reconstruction. In this section we describe the existing approaches for detecting reticulate events at the population as well as species levels.

5.1.1 Recombination

Detecting recombination is a major topic of study in population genetics, with a commensurate number of publications. Studies of specific systems abound—any literature search

using the keyword "recombination" will immediately bring up hundreds and any text on genetics will at least discuss the topic. Mostly, of course, such recombinations are meiotic in nature. In phylogenetic work, detecting recombinations (this time from a variety of sources) is at the heart of many approaches to the reconstruction of ancestral genomes or lines of descent [89, 98, 99, 106, 163, 165, 247, 261, 283]. Wiuf *et al.* [279] and Posada and Crandall [199] have studied the accuracy of methods for detecting recombination from a collection of DNA sequences; their papers contain and a wealth of references. Detecting the presence of recombination is but the first step; characterizing the recombinations that did take place is the goal. An intermediate goal along this path is to determine which recombination events might have taken place, as is done in many studies and implemented in programs such as SplitsTree [114] and NeighborNet [29]. A further step could determine the number of recombination events that took place [252], a goal that programs such as T-Rex [151] and the reconstruction algorithm of Hallett and Lagergren [3] attempt to reach. The last step is to produce a collection of recombination events that best explain the observed data, i.e., to produce one or more evolutionary networks that optimize some criterion (perhaps a generalization of a criterion used in tree reconstruction, such as minimum evolution, parsimony or maximum likelihood); few researchers have tackled that problem directly and none of the existing programs addresses it.

5.1.2 Horizontal Gene Transfer

Once again, the first order of business is to detect horizontal transfers. The goal of many biological studies has been to identify those genes that were acquired by the organism through horizontal transfers rather than inherited from its ancestors. In one of the first papers on the topic, Medigue [166] proposed the use of multivariate analysis of codon usage to identify such genes; since then various authors have proposed other intrinsic methods, such as using GC content, particularly in the third position of codons (e.g., [136]). On the basis of such approaches, a database of genes acquired by horizontal transfer in prokaryotes has been established at `www.fut.es/~debb/HGT/`. An advantage of intrinsic approaches is their ability to identify and eliminate genes that do not obey a tree-like process of evolution and thus could prevent classical phylogenetic methods from reconstructing a good tree. With the advent of whole-genome sequencing, more powerful intrinsic methods become possible, such as the location of suspect genes with each genome: such locations tend to be preserved through lineages, but a transfer event can place the new gene in a more or less random location. Thus a good look at the neighbors of a gene in the prokaryotic genome often enables biologists to identify horizontal transfers. However, differential selection pressure, uneven evolutionary rates, and biased sampling can all give rise to false identification of HGT [67].

Non-intrinsic approaches bring us back to phylogenetic reconstruction. Here, the idea is to use phylogenetic reconstructions to identify discrepancies that could tag transfer events. An old question in phylogenetic reconstruction has been "to combine or not to combine?"—that is, given DNA sequences for several genes, are we better off concatenating the sequences or analyzing each set separately [31, 40, 47, 53, 109, 185, 277]? The common sense conclusion that many genes inherited through lineal descent would

override the confusing signal generated by a few genes acquired through horizontal transfer appears wrong [25, 266]. Of course, one must first resolve the old problem of gene trees vs. species trees: discrepancies between the trees derived from different genes do not necessarily indicate reticulate evolution, but may simply testify to the incongruent evolution of two or more genes, all within a valid, tree-shaped evolution of the species. (For discussions of and algorithms for the gene tree/species tree, or reconciliation, problem and some of its related problems, such as distinguishing orthologs from paralogs, see [12, 70, 148, 150, 190, 191, 194, 230, 256].) Distinguishing between the two is difficult in the absence of additional information.

With whole-genome sequencing, such information becomes available. Huynen and Bork [115] advocate two types of data: the fraction of shared orthologs and gene synteny. Synteny (the conservation of genes on the same chromosome) is of course not widely applicable with prokaryotes, but its logical extension, conservation of gene order, definitely is—and Huynen and Bork proposed to measure the fraction of conversed adjacencies, a notion that had been introduced earlier by Sankoff in a series of papers defining breakpoints (adjacencies that are not conserved) and their uses [22, 237]. Orthologs are a phylogenetic notion: two homologous genes are orthologs if they are the product of speciation from a common ancestor; in contrast, two homologous genes are paralogs if they are the product of duplication. Thus determining orthologs can be difficult. Daubin [50] combined orthology search and information from the DNA sequences themselves to improve the detection of horizontal transfers.

From a computing point of view, the problem can still be formulated as pure network reconstruction. Direct approaches include those of Hallett and Lagergren [3, 95] and Boc and Makarenkov [23].

5.1.3 Hybrid speciation

Funk [79] and McDade [161] discuss the implications of hybridization for phylogenetic studies. Rieseberg and his colleagues have published numerous papers with case studies, discussions, and methodologies, all addressed at diploid hybrid speciation [69, 209, 210, 211, 213, 214, 215, 218, 219, 223, 224, 225]. Detecting hybrid speciation can be trivial: when the ploidy level (i.e., number of complete sets of chromosomes) changes, clearly hybrid speciation has occurred [211]. In other cases, the detection of hybrid speciation can be more difficult, especially when the parent species have similar sequences. As in the case of horizontal gene transfer, the true evolutionary history requires the network model. The two types of reticulate evolution—HGT and hybrid speciation—differ in significant ways. In HGT, the number of lineages does not change, but hybrid speciation directly creates a new lineage. In addition, under HGT the proportion of the donor genome to the recipient genome is relatively small, whereas in hybrid speciation is more nearly equal.

As discussed below, there are a small number of methods that attempt to detect and reconstruct hybrid speciation events [17, 29, 114, 238, 285], but none are entirely satisfactory, especially at reconstruction. In general, the methods produce an unacceptable number of false positives (edges that appear in the reconstruction, but are not part of the true network). The problems most likely arise because the methods tend to use combined data and

because they lack sufficient biological rationale.

5.2 Methods for Reconstructing Networks

Broadly speaking, four approaches to phylogenetic reconstruction in the presence of non-tree events have been developed. Since these have not been sufficiently studied in simulation studies, we do not yet know how well they perform, or whether these approaches are promising. Our discussion of the current network reconstruction methods is, therefore, quite brief.

1. With prokaryotic organisms, where the non-tree event is horizontal gene transfer (HGT), one common approach is to identify (by whatever means) the genes acquired through HGT, remove them, then reconstruct a tree based on the remaining genes. Most published bacterial phylogenies are trees and were either limited to one or a few genes for data, or use the whole genome, but only after eliminating genes identified as the product of HGT.
2. One can build a tree, then begin adding non-tree edges to turn it into a network, using a greedy approach to optimize some cost criterion. This is the approach used by Hallett and Lagergren [3, 95] (for horizontal transfers only), by Clement *et al.* [42], and by Makarenkov and his colleagues [151, 152].
3. Build many trees (perhaps using different subsets of the data) and attempt to reconcile them; where reconciliation fails, the conflict might be explained by a reticulation event. This is the basic idea behind median networks [17, 18, 19], as well as behind the molecular-variance parsimony approach of Excoffier *et al.* [71].
4. Characterize in advance any incompatibilities in the data (based on a distance matrix) and provide a collection of the possible resolutions through reticulation—leaving the biologist to choose which resolution is preferable. This approach is used in the splits-based methods [16, 29, 107, 114].

The first approach of course does not build a network; however, in groups of prokaryotes, where gene transfer is ubiquitous and there may not even exist a true "core" of genes, it has the advantage of representing the structured part of the information, with the understanding that nodes in the tree can be connected by a multitude of non-tree edges corresponding to HGTs. The last approach does not build or even propose a specific network, but presents all consistent choices—a potential problem when the number of choices is large.

5.3 Simulation Tools for Phylogenetic Networks

Software tools for generating random phylogenetic networks and simulating sequence evolution down phylogenetic networks, have been developed for the case of hybrid speciation [178]. These tools are adaptations of those used for the simulation of tree evolution, but, whereas tree evolution can proceed independently on the various branches, network evolution perforce reconnects various paths. Hence, whereas tree simulations can first create a

tree topology and then evolve sequences down that tree, network simulations must evolve the sequences as they create the topology, since the choice of hybridization events (and, to a lesser extent, of horizontal transfer events) depends on the genomic distances between the two organisms involved in the hybridization or transfer. To our knowledge, no publically available software exists for generating random networks and simulating sequence evolution in the case of horizontal transfer nor for combining these reticulation events with recombination.

5.4 Measuring Error in Network Reconstruction

Performance studies also need to be able to measure the error (distance) between the "true" phylogeny and the reconstructed one. Such a measure should be symmetric, and should be zero only when the two phylogenies are the same. Such measures are commonplace for trees; the most commonly used is the Robinson-Foulds distance [227] (see Section 2.4), but quartet-based metrics also provide these properties. However, such measures between phylogenetic networks seem much harder to obtain; our group has developed some measures,but the guarantees we seek (in particular, being zero if and only if the two networks are identical) only hold under special conditions. Clearly, additional work needs to be done in this regard.

We briefly describe two approaches to this problem. We want a measure $m(N_1, N_2)$ of the distance between two networks N_1 and N_2, such that m is symmetric and nonnegative, and satisfies the following three conditions:

C1 If N_1 and N_2 are two trees, then $m(N_1, N_2) = RF(N_1, N_2)$, where $RF(N_1, N_2)$ denotes the Robinson-Foulds distance between trees.

C2 If N_1 and N_2 are isomorphic (with respect to the leaf labels), then $m(N_1, N_2) = 0$.

C3 If $m(N_1, N_2) = 0$, then N_1 and N_2 are isomorphic.

Splits-based measure One appoach that has been used both by us and by Bryant [29] is a direct extension of the Robinson-Foulds (RF) metric on trees. Recall that the RF distance is just half the size of the symmetric difference between the character encodings of the two trees, and so is naturally symmetric and zero if and only if the two trees are identical. Let N be a network, and let S be the set of leaves of N. Let $\mathcal{T}$ be the collection of trees, leaf -labelled by the set S of taxa, contained within the network N. Let $\mathcal{C}(N) = \cup_{T \in \mathcal{T}} C(T)$ (i.e., the set containing all the bipartitions of the trees in $\mathcal{T}$). Given two networks N_1 and N_2 we can define the "distance" between them to be:

$$d(N_1, N_2) = \frac{|C(N_1) \triangle C(N_2)|}{2}.$$

This distance satisfies conditions C1 and C2 above, but not, unfortunately, C3, except in special cases.

Tripartition measure Another proposed measure is based on the tripartition that is induced by each edge of the network. Let N be a phylogenetic network, leaf-labeled by set S, and let $e = (u, v)$ be an edge of N. Note then that an edge e induces a tripartition of S, defined by the sets

- $A(e) = \{s \in S \mid s \text{ is reachable from the root of } N \text{ only via } v\}$.
- $B(e) = \{s \in S \mid s \text{ is reachable from the root of } N \text{ via at least one path passing through } v \text{ and one path not passing through } v\}$.
- $C(e) = \{s \in S \mid s \text{ is not reachable from the root of } N \text{ via } v\}$.

We denote by $\theta(e)$ the tripartition of S induced by edge e. These three sets $A(e)$, $B(e)$ and $C(e)$ are weighted; the weight of an element s in any of the three sets is the maximum number of network nodes on a path from v to s (including v and/or s, in case one or both are network nodes). Two weighted sets S_1 and S_2 are equivalent, denoted by $S_1 \equiv S_2$, whenever they contain the same elements and each element has the same weight in both sets. Two tripartitions $\theta(e_1)$ and $\theta(e_2)$ are equivalent, denoted by $\theta(e_1) \equiv \theta(e_2)$, whenever we have $A(e_1) \equiv A(e_2)$, $B(e_1) \equiv B(e_2)$, and $C(e_1) \equiv C(e_2)$.

Let x be a network node whose two parents are x_1 and x_2. The *reticulation scenario* of x, denoted $RS(x)$, is the set $\{X_1, X_2\}$, where X_i is the set of leaves under x_i (that is, if x_i is the head of edge e, then $X_i = A(e) \cup B(e)$). Intuitively, a reticulation scenario of a network node x denotes the two groups of taxa whose common ancestors were involved in the reticulation event.

We can now define compatibility between two edges as follows. Let $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$ be two edges. Then, e_1 and e_2 are compatible, denoted by $e_1 \equiv e_2$, if and only if

- either both are tree edges and $\theta(e_1) \equiv \theta(e_2)$, or
- both are network edges, $\theta(e_1) \equiv \theta(e_2)$, and $RS(v_1) = RS(v_2)$.

Finally, we can define the *false negative rate* (FN) and *false positive rate* (FP) between two networks N_1 and N_2 in the usual way: they are the percentages of edges present in one network that have no compatible edge (as per the definition above) in the other.

This tripartition metric satisfies all three conditions, C1, C2, and C3.

6 Summary

Reticulate evolution, due to factors such as horizontal gene transfer, recombination, and speciating hybridization, confound current approaches to phylogenetic analysis, and an accurate estimation of evolutionary histories will require the development of novel simulation tools, reconstruction methods, and mathematical theory. We conjecture that successful approaches in this area will combine population genetics and phylogenetics, and will lead to interesting questions in many technical areas, including statistical inference, molecular phylogenetics, and computer science.

7 Acknowledgments

We thank the organizers of PSB 2004 for the opportunity to present this tutorial. Our work is supported by the National Science Foundation under grants ACI 00-81404 (Moret), DEB 01-20709 (Linder, Moret, and Warnow), EF 03-31453 (Warnow), EF 03-31654 (Moret), EIA 02-03584 (Moret), EIA 01-13095 (Moret), EIA 01-13654 (Warnow), EIA 01-21377 (Moret), and EIA 01-21680 (Linder and Warnow), by a grant from IBM Corporation (Moret), by the David and Lucile Packard Foundation (Warnow), and by a Radcliffe Institute for Advanced Study Fellowship (Warnow).

8 Software Related to Network Reconstruction

- **Pyramids**
  - `http://genome.genetique.uvsq.fr/Pyramids/`
  - Authors: J.C. Aude *et al.*
  - This program takes a distance matrix as an input and uses agglomerative algorithms to compute clusters. It allows for overlapping clusters, which can be used to represent reticulation events. Reticulation events may be placed among terminal nodes that are sister taxa.
  - Described in [57]
  - Platforms: Executables for Windows and Unix
- **Spectrum**
  - `http://taxonomy.zoology.gla.ac.uk/~mac/spectrum/spectrum.html`
  - Authors: Impkin Software (a division of Impkin Incorporated)
  - This program takes a set of sequences in the NEXUS format as an input and visualizes phylogenetic information without forcing it into a tree, thus avoiding the difficulty of choosing which is the "best" method for tree reconstruction and the whole issue of whether the data is tree-like.
  - Platforms: Executables for Mac and Windows are available
- **Arlequin**
  - `http://lgb.unige.ch/arlequin/`
  - Authors: Laurent Excoffier
  - This program supports population genetics analysis, such as estimation of gene frequencies, testing of linkage disequilibrium, and analysis of diversity between populations. It can also compute a Minimum Spanning Tree network.
  - Platforms: Mac, Windows, and Linux
- **PAL**

  - http://www.stat.uni-muenchen.de/∼strimmer/pal-project/
  - Authors: Alexei Drummond and Korbinian Strimmer
  - A collection of Java classes for use in molecular phylogenetics. Among the tasks that have been implemented using this collection is computing the maximum likelihood of phylogenetic networks.
  - Platforms: JAVA source code

- **T-REX**
  - http://www.fas.umontreal.ca/biol/casgrain/en/labo/t-rex/
  - Authors: Vladimir Makarenkov
  - This program takes a distance matrix as an input, infers a tree, and then adds non-tree edges, thus creating a network, so as to minimize a certain least-squares loss function.
  - Described in [151]
  - Platforms: Windows, DOS, and Mac binaries, as well as C++ source code

- **PLATO**
  - http://evolve.zoo.ox.ac.uk/Plato/main.html
  - Authors: Nick Grassly
  - This program takes sequential PHYLIP-style DNA sequences and their maximum likelihood phylogeny; using a likelihood approach with sliding window analysis and Monte Carlo simulation of the null distribution, it then detects anomalously evolving regions in the DNA sequences and assesses their significance. This may lead to the detection of, for example, recombination, gene conversion, or convergence; it may also reveal variable selective pressures along the gene sequences.
  - Described in [87].
  - Platforms: Mac (including PowerMacs) and source code for Unix systems

- **Bootscanning Package**
  - by anonymous ftp from http://www.ktl.fi in directory /hiv/mirrors/pub/programs
  - Authors: Mika Salminen and Wayne Cobb
  - A series of shell scripts and programs that analyze DNA sequences for evidence of recombination. The code breaks the sequence into separate pieces that are analyzed for the bootstrap support of various groups; the code looks for evidence of significant conflict among trees for different parts of the sequence.
  - Platforms: Sun executables

- **TOPAL**
  - http://www.bioss.sari.ac.uk/ frank/Genetics

  - Authors: Grainne McGuire
  - A collection of scripts and code that checks for evidence of past recombination events by looking for changes in the inferred phylogenetic tree TOPology between adjacent regions of a multiple sequence ALignment. The method detects recombinations by sliding a window along a sequence alignment and measuring the discrepancy between the trees suggested by the first and second halves of the window, using distance matrix methods.
  - Described in [162].
  - Platforms: Unix Bourne shell scripts and C code

- **reticulate**
  - `http://jcsmr.anu.edu.au/dmm/humgen/ingrid/reticulate.htm`
  - Authors: Ingrid Jakobsen and Simon Easteal
  - A compatibility matrix program for DNA sequences that includes tests for evidence of reticulate evolution (such as recombination). The program computes and displays a pairwise compatibility matrix for all pairs of sites and provides statistics on compatible pairs.
  - Described in [119]
  - Platforms: C source code for Unix and X Windows

- **RecPars**
  - `ftp.daimi.aau.dk` in directory `pub/empl/kfisker/programs/RecPars`
  - Authors: Kim Fisker
  - This program runs a parsimony analysis of DNA sequences and tries to find the best phylogenies for different regions of the sequences; different phylogenies lead it to postulate a recombination event between the respective segments.
  - Described in [99]
  - Platforms: C source code for Unix

- **Partimatrix**
  - `http://jcsmr.anu.edu.au/dmm/humgen/ingrid/partimatrix.htm`
  - Authors: Ingrid Jakobsen, Susan Wilson, and Simon Easteal
  - This program computes a "partition matrix" from aligned DNA sequence data. It finds partitions of the sequences into two groups and presents a matrix that describes conflicts and agreements among these partitions.
  - Described in [120]
  - Platforms: C source code for Unix systems with X Windows

- **Homoplasy test**
  - `http://www.biols.susx.ac.uk/Home/John_Maynard_Smith/`

  - Authors: John Maynard Smith and Noel Smith
  - This program implements the authors' homoplasy test for recombination in sequences.
  - Described in [247]
  - Platforms: QBASIC for DOS and must be run using QBASIC

- **LARD**
  - http://evolve.zoo.ox.ac.uk/software/Lard/Lard.html
  - Authors: Andrew Rambaut
  - This program detects the presence of recombination in a set of sequences. LARD looks at the set of sequences to discover which are the most plausible parents of a potentially recombinant sequence; at each possible breakpoint position, it runs a likelihood ratio test to check whether the three-species trees differ on the two sides of the breakpoint.
  - Described in [106]
  - Platforms: C source code and as a Macintosh executable

- **Network**
  - http://www.fluxus-engineering.com/sharenet.htm
  - Authors: Arne Röhl, Peter Forster, and Hans-Jürgen Bandelt
  - This program infers networks (which have more connections than trees) from non-recombining DNA, STR, amino acid, and RFLP data. The networks are either reduced median networks nor median-joining networks.
  - Described in [18, 17]
  - Platforms: DOS executable

- **TCS**
  - http://bioag.byu.edu/zoology/crandall_lab/tcs.htm
  - Authors: Mark Clement and David Posada
  - This program estimates gene genealogies within a population, using a method that connects existing haplotypes in a minimum spanning tree (essentially a parsimony method).
  - Described in [268, 42]
  - Platforms: Java executables

- **SplitsTree**
  - http://www-ab.informatik.uni-tuebingen.de/software/splits/
  - Authors: D. Huson
  - This program outputs a network as a graphical representation of the incompatibilities in the dataset.

  - Described in [114]
  - Platforms: Unix and Windows

- **NeighborNet**
  - http://www.mcb.mcgill.ca/~bryant/NeighborNet/
  - Authors: D. Bryant and V. Moulton
  - This program constructs a network that is a graphical representation of the compatibilities in the dataset.
  - Described in [29]
  - Platforms: Unix

- **Horizontal Transfer Program**
  - http://www.cs.mcgill.ca/~laddar/lattrans/
  - Authors: L. Addario Berry, M. Hallett, and J. Lagergren
  - This program outputs a tree and a list of the horizontal gene transfer scenarios (i.e., a list of extra nontree edges that correspond to transfer events).
  - Described in [3]
  - Platforms: Java code

References

[1] R.J. Abbott. Plant invasions, interspecific hybridization and the evolution of new plant taxa. *Trends in Ecol. and Evol.*, 7:401–405, 1992.

[2] K. Adams, D. Daley, Y.-L. Qui, J. Whelan, and J.D. Palmer. Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. *Nature*, 408:354–357, 2000.

[3] L. Addario-Berry, M.T. Hallett, and J. Lagergren. Towards identifying lateral gene transfer events. In *Proc. 8th Pacific Symp. on Biocomputing (PSB03)*, pages 279–290, 2003.

[4] J. Alroy. Continuous track analysis: A new phylogenetic and biogeographic method. *Syst. Biol.*, 44(2):152–178, 1995.

[5] A. Anderberg and A. Tehler. Consensus trees, a necessity in taxonomic practice. *Cladistics*, 6:399–402, 1990.

[6] E. Anderson. *Introgressive Hybridization*. John Wiley and Sons, 1949.

[7] J.O. Andersson, W.F. Doolittle, and C.L. Nesbo. Genomics—are there bugs in our genome? *Science*, 292(5523):1848–1850, 2001.

[8] M.L. Arnold. Natural hybridization as an evolutionary process. *Ann. Rev. Ecol. Syst.*, 23:237–261, 1992.

[9] M.L. Arnold. Natural hybridization and Louisiana irises. *Bioscience*, 44(3):141–147, 1994.

[10] M.L. Arnold. *Natural Hybridization and Evolution*. Oxford U. Press, 1997.

[11] M.L. Arnold, C.M. Buckner, and J.J. Robinson. Pollen-mediated introgression and hybrid speciation in Louisiana irises. *Proc. Nat'l Acad. Sci., USA*, 88(4):1398–1402, 1991.

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

[13] W.R. Atchley and W. Fitch. Gene trees and the origin of inbred strains of mice. *Science*, 254:554–558, 1991.

[14] W.R. Atchley and W. Fitch. Genetic affinities of inbred mouse strains of uncertain origin. *Mol. Biol. Evol.*, 10:1150–1169, 1993.

[15] D.A. Bader, B.M.E. Moret, and L. Vawter. Industrial applications of high-performance computing for phylogeny reconstruction. In H.J. Siegel, editor, *Proc. SPIE Commercial Applications for High-Performance Computing (SPIE01)*, volume 4528, pages 159–168, 2001.

[16] H.J. Bandelt and A.W.M. Dress. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phyl. Evol.*, 1:242–252, 1992.

[17] H.J. Bandelt, P. Forster, and A. Roehl. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.*, 16(1):37–48, 1999.

[18] H.J. Bandelt, P. Forster, B.C. Sykes, and M.B. Richards. Mitochondrial portraits of human populations using median networks. *Genetics*, 141:743–753, 1995.

[19] H.J. Bandelt, V. Macaulay, and M. Richards. Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol. Phyl. Evol.*, 16:8–28, 2000.

[20] M. Barrett, M.J. Donoghue, and E. Sober. Against consensus. *Syst. Zool.*, 40(4):486–493, 1991.

[21] N.H. Barton. The role of hybridization in evolution. *Molecular Ecology*, 10(3):551–568, 2001.

[22] M. Blanchette, G. Bourque, and D. Sankoff. Breakpoint phylogenies. In S. Miyano and T. Takagi, editors, *Genome Informatics*, pages 25–34. Univ. Academy Press, Tokyo, 1997.

[23] A. Boc and V. Makarenkov. New efficient algorithm for detection of horizontal gene transfer events. In *Proc. 3rd Int'l Workshop Algorithms in Bioinformatics (WABI03)*, volume 2812, pages 190–201. Springer-Verlag, 2003.

[24] P.L. Bollyky, A. Rambaut, P.H. Harvey, and E.C. Holmes. Recombination between sequences of hepatitis B virus from different genotypes. *J. Mol. Evol.*, 42:97–102, 1996.

[25] J.R. Brown, C.J. Douady, M.J. Italia, W.E Marshall, and M.J. Stanhope. Universal trees based on large combined protein sequence data sets. *Nat. Genet.*, 28:281–285, 2001.

[26] J.R. Brown and P.V. Warren. Antibiotic discovery: Is it in the genes? *Drug Discovery Today*, 3:564–566, 1998.

[27] W.J. Bruno and A.L. Halpern. Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.*, 16:564–566, 1999.

[28] D. Bryant, D. Huson, T. Kloepper, and K. Nieselt-Struwe. Phylogenetic analysis of recombinant sequences. In *Proc. 3rd Int'l Workshop Algorithms in Bioinformatics (WABI03)*, volume 2812, pages 271–286. Springer-Verlag, 2003.

[29] D. Bryant and V. Moulton. NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. In *Proc. 2nd Int'l Workshop Algorithms in Bioinformatics (WABI02)*, volume 2452 of *Lecture Notes in Computer Science*, pages 375–391. Springer-Verlag, 2002.

[30] C.A. Buerkle, R.J. Morris, M.A. Asmussen, and L.H. Rieseberg. The likelihood of homoploid hybrid speciation. *Heredity*, 84(4):441–451, 2000.

[31] J.J. Bull, J.P. Huelsenbeck, C.W. Cunningham, D. Swofford, and P. Waddell. Partitioning and combining data in phylogenetic analysis. *Syst. Biol.*, 42(3):384–397, 1993.

[32] L. Bullini. Origin and evolution of animal hybrid species. *Trends in Ecol. and Evol.*, 9(11):422–426, 1994.

[33] A.M. Campbell. Lateral gene transfer in prokaryotes. *Theoretical Population Biology*, 57(2):71–77, 2000.

[34] C. Canchaya, G. Fournous, S. Chibani-Chennoufi, M.L. Dillmann, and H. Brussow. Phage as agents of lateral gene transfer. *Current Opinion in Microbiology*, 6(4):417–424, 2003.

[35] S.B. Carroll, J.K. Grenier, and S.D. Weatherbee. *From DNA to Diversity*. Blackwell Science, 2001.

[36] J.K. Chambers and L.E. McDonald *et al.* A G protein-coupled receptor for UDP-glucose. *J. Biol. Chem.*, 275(15):10767–10771, 2000.

[37] J. Chang. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Math. Biosciences*, 134:189–215, 1996.

[38] M.A. Charleston. Jungles: A new solution to the host/parasite phylogeny reconciliation problem. *Math. Biosciences*, 149:191–223, 1998.

[39] D. Charlesworth. Hybrid speciation—evolution under the microscope. *Current Biology*, 5(8):835–836, 1995.

[40] P.T. Chippindale and J.J. Wiens. Weighting, partitioning, and combining characters in phylogenetic analysis. *Syst. Biol.*, 43(2), 1994.

[41] M.T. Clegg, B.S. Gaut, M.R. Duvall, and J. Davis. Inferring plant evolutionary history from molecular data. *New Zealand Journal of Botany*, 31(3):307–316, 1993.

[42] M. Clement, D. Posada, and K. Crandall. TCS: a computer program to estimate gene genealogies. *Mol. Ecol.*, 9:1657–1660, 2000.

[43] H.P. Comes and R.J. Abbott. Reticulate evolution in the Mediterranean species complex of Senecio sect Senecio: Uniting phylogenetic and population-level approaches. In P. Hollingsworth, R. Bateman, and R. Gornall, editors, *Molecular Systematics and Plant Evolution*, pages 171–198. Taylor and Francis, 1999.

[44] J.F. Crow and M. Kimura. *An Introduction to Population Genetics Theory*. Harper and Row, 1970. Reprinted by Burgess Int'l.

[45] M.B. Cruzan. Genetic markers in plant evolutionary ecology. *Ecology*, 79(2):400–412, 1998.

[46] M.P. Cummings. Transmission patterns of eukaryotic transposable elements: arguments for and against horizontal transfer. *Trends in Ecol. and Evol.*, 9:141–145, 1994.

[47] C.W. Cunningham. Can three incongruence tests predict when data should be combined? *Mol. Biol. Evol.*, 14:733–740, 1997.

[48] C.W. Cunningham. Is congruence between data partitions a reliable predictor of phylogenetic accuracy? Empirically testing an iterative procedure for choosing among phylogenetic methods. *Syst. Biol.*, 46:464–478, 1997.

[49] P. Darlu and G. Lecointre. When does the incongruence length difference test fail? *Mol. Biol. Evol.*, 19:432–437, 2002.

[50] V. Daubin, M. Gouy, and G. Perriere. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.*, 12:1080–1090, 2002.

[51] A.P. de Koning, F.S. Brinkman, S.J. Jones, and P.J. Keeling. Lateral gene transfer and metabolic adaptation in the human parasite Trichomonas vaginalis. *Mol. Biol. Evol.*, 17(11):1769–1773, 2000.

[52] F. de la Cruz and J. Davies. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.*, 8:128–133, 2000.

[53] A. de Queiroz, M.J. Donoghue, and J. Kim. Separate versus combined analysis of phylogenetic evidence. *Annu. Rev. Ecol. Syst.*, 25:657–681, 1995.

[54] R. de Rosa, J. K. Grenier, T. Andreeva, C. E. Cook, A. Adoutte, M. Akam, S. B. Carroll, and G. Balavoine. Hox genes in brachipods and priapulids and protostome evolution. *Nature*, 399:772–776, 1999.

[55] C.F. Delwiche and J.D. Palmer. Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. *Mol. Biol. Evol.*, 13(6):873–882, 1996.

[56] C.F. Delwiche and J.D. Palmer. The origin of plastids and their spread via secondary symbiosis. *Plant Systematics and Evolution*, pages 53–86, 1997.

[57] E. Diday and P. Bertrand. An extension of hierarchical clustering: the pyramidal representation. In E.S. Gelsema and L.N. Kanal, editors, *Pattern Recognition in Practice II*, pages 411–424. Elsevier Science Publ., 1986.

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

[59] W.F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284:2124–2129, 1999.

[60] C.G. Dowson, V. Barcus, S. King, P. Pickerill, A. Whatmore, and M. Yeo. Horizontal gene transfer and the evolution of resistance and virulence determinants in streptococcus. *Journal of Applied Microbiology*, 83:S42–S51, 1997.

[61] M. Dowton and A.D. Austin. Increased congruence does not necessarily indicate increased phylogenetic accuracy—the behavior of the incongruence length difference test in mixed-model analyses. *Syst. Biol.*, 51:19–31, 2002.

[62] J.J. Doyle. Gene trees and species trees: Molecular systematics as one-character taxonomy. *Syst. Bot.*, 17:144–163, 1992.

[63] J.J. Doyle, J.L. Doyle, J.A. Ballenger, and J.D. Palmer. The distribution and phylogenetic significance of a 50 kb chloroplast dna inversion in the flowering plant family Leguminosae. *Mol. Phyl. Evol.*, 5:429–438, 1996.

[64] J.J. Doyle, D.E. Soltis, and P.S. Soltis. An intergeneric hybrid in the saxifragaceae: Evidence from ribosomal RNA genes. *American Journal of Botany*, 72(9):1388–1391, 1985.

[65] S. Dumolin-Lapgue, A. Kremer, and R.J. Pettit. Are chloroplast and mitochondrial DNA variation species independent in oaks. *Evolution*, 53(5):1406–1413, 1999.

[66] C. Dutta and A. Pan. Horizontal gene transfer and bacterial diversity. *Journal of Biosciences*, 27(1):27–33, 2002.

[67] J.A. Eisen. Horizontal gene transfer among microbial genomes: New insights from complete genome analysis. *Curr Opin Genet Dev.*, 10(6):606–611, 2000.

[68] N.C. Ellstrand and C.A. Hoffman. Hybridization as an avenue of escape for engineered genes. *BioScience*, 40(6):438–442, 1990.

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

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

[71] L. Excoffier, P.E. Smouse, and J.M. Quattro. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131:479–491, 1992.

[72] M. Fellows, M.T. Hallett, and U. Stege. On the multiple gene duplication problem. In *Proc. 9th Int'l. Symp. Alg. and Comp. ISAAC98*, volume 1533 of *Lecture Notes in Computer Science*, pages 347–356. Springer-Verlag, 1998.

[73] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.*, 27:401–410, 1978.

[74] J. Felsenstein. Phylogenies and the comparative method. *Amer. Nat.*, 125:1–15, 1985.

[75] J. Felsenstein. The troubled growth of statistical phylogenetics. *Syst. Biol.*, 50(4):465–467, 2001.

[76] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, MA, 2003.

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

[78] L.R. Foulds and R.L. Graham. The steiner tree problem in phylogeny is np-complete. *Advances in Appl. Math.*, 3:43–49, 1982.

[79] V.A. Funk. Phylogenetic patterns and hybridization. *Ann. Mo. Bot. Gard.*, 72:681–715, 1985.

[80] D.J. Futuyma. *Evolutionary Biology*. Sinauer Assoc., Sunderland, MA, 1998.

[81] M.Y. Galperin and E.V. Koonin. Comparative genome analysis. *Methods Biochem. Anal.*, 43:359–392, 2001.

[82] J.P. Gogarten, W.F. Doolittle, and J.G. Lawrence. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.*, 19(12):2226–2238, 2002.

[83] M. Goodman, J. Czelusniak, G. Moore, E. Romero-Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.*, 28:132–163, 1997.

[84] D.E. Graham, R. Overbeek, G.J. Olsen, and C.R. Woese. An archeal genomic signature. *Proc. Nat'l Acad. Sci., USA*, 97:3304–3308, 2000.

[85] S.W. Graham, J.R. Kohn, B.R. Morton, J.E. Eckenwalder, and S.C.H. Barrett. Phylogenetic congruence and discordance among one morphological and three molecular data sets from Pontederiaceae. *Syst. Biol.*, 47(4):545–567, 1998.

[86] V. Grant. *Plant Speciation*. Columbia University Press, New York, 1971.

[87] N.C. Grassly and E.C. Holmes. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.*, 14:239–247, 1997.

[88] S. Gribaldo and H. Philippe. Ancient phylogenetic relationships. *Theoretical Population Biology*, 61(4):391–408, 2002.

[89] R.C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, 3:479–502, 1996.

[90] X. Gu. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.*, 18(4):453–464, 2001.

[91] R.S. Gupta. Evolutionary relationships among photosynthetic bacteria. *Photosynthesis Research*, 76(1–3):173–183, 2003.

[92] R.S. Gupta and E. Griffiths. Critical issues in bacterial phylogeny. *Theoretical Population Biology*, 61(4):423–434, 2002.

[93] P. Halbur, M.A. Lum, X. Meng, I. Morozov, and P.S. Paul. New porcine reproductive and respiratory syndrome virus DNA and proteins encoded by open reading frames of an Iowa strain of the virus are used in vaccines against PRRSV in pigs, 1994. Patent filing WO9606619-A1.

[94] M.T. Hallett and J. Lagergren. New algorithms for the duplication-loss model. In *Proc. 4th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB00)*, pages 138–146, New York, 2000. ACM Press.

[95] M.T. Hallett and J. Lagergren. Efficient algorithms for lateral gene transfer problems. In *Proc. 5th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB01)*, pages 149–156, New York, 2001. ACM Press.

[96] A.L. Halpern and W.J. Bruno. Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. *Mol. Biol. Evol.*, 15:910–917, 1998.

[97] J.R. Harlan. Evolutionary dynamics of plant domestication. *Japanese Journal of Genetics*, 44(1):337–343, 1969.

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

[99] J. Hein. A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.*, 36:396–405, 1993.

[100] D.M. Hillis. Molecular versus morphological approaches to systematics. *Annu. Rev. Ecol. Syst.*, 18:23–42, 1987.

[101] D.M. Hillis. Primer: Phylogenetic analysis. *Current Biology*, 7:R129–R131, 1997.

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

[103] D.M. Hillis and J.P. Huelsenbeck. To tree the truth: Biological and numerical simulations of phylogeny. In D.M. Fambrough, editor, *Molecular Evolution of Physiological Processes*, pages 55–67. Rockefeller University Press, 1994.

[104] D.M. Hillis and J.P. Huelsenbeck. Assessing molecular phylogenies. *Science*, 267:255–256, 1995.

[105] D.M. Hillis, B.K. Mable, and C. Moritz. *Molecular Systematics*. Sinauer Assoc., Sunderland, Mass., 1996.

[106] E.C. Holmes, M. Worobey, and A. Rambaut. Phylogenetic evidence for recombination in dengue virus. *Mol. Biol. Evol.*, 16:405–409, 1999.

[107] K.T. Huber, E.E. Watson, and M.D. Hendy. An algorithm for constructing local regions in a phylogenetic network. *Mol. Phyl. Evol.*, 19(1):1–8, 2001.

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

[109] J.P. Huelsenbeck, J.J. Bull, and C.W. Cunningham. Combining data in phylogenetic analysis. *Trends in Ecol. and Evol.*, 11(4):151–157, 1996.

[110] J.P. Huelsenbeck, B. Rannala, and B. Larget. A Bayesian framework for the analysis of cospeciation. *Evol.*, 54(2):353–364, 2000.

[111] J.P. Huelsenbeck, B. Rannala, and Z. Yang. Statistical tests of host-parasite cospeciation. *Evol.*, 51(2):410–419, 1997.

[112] J.P. Huelsenbeck, D. Swofford, C.W. Cunningham, J.J. Bull, and P.W. Waddell. Is character weighting a panacea for the problem of data heterogeneity in phylogenetic analysis? *Syst. Biol.*, 43(2):288–291, 1994.

[113] D. H. Huson, S. Nettles, K. Rice, and T. Warnow. Hybrid tree reconstruction methods. In *Proc. 2nd Worshop on Algorithm Engineering WAE98*, Saarbrücken, 1998.

[114] D.H. Huson. SplitsTree: a program for analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73, 1998.

[115] M.A. Huynen and P. Bork. Measuring genome evolution. *Proc. Nat'l Acad. Sci., USA*, 95:5849–5856, 1998.

[116] R. Jain, M.C. Rivera, and J.A. Lake. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Nat'l Acad. Sci., USA*, 86:3801–3806, 1999.

[117] R. Jain, M.C. Rivera, J.E. Moore, and J.A. Lake. Horizontal gene transfer in microbial genome evolution. *Theoretical Population Biology*, 61(4):489–495, 2002.

[118] R. Jain, M.C. Rivera, J.E. Moore, and J.A. Lake. Horizontal gene transfer accelerates genome innovation and evolution. *Mol. Biol. Evol.*, 20(10):1598–1602, 2003.

[119] I.B. Jakobsen and S. Eastel. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Bioinformatics*, 12:291–295, 1996.

[120] I.B. Jakobsen, S.R. Wilson, and S. Eastel. The partition matrix: Exploring variable phylogenetic signals along nucleotide sequence alignments. *Mol. Biol. Evol.*, 14(5):474–484, 1997.

[121] R.K. Jansen and J.D. Palmer. A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). *Proc. Nat'l Acad. Sci., USA*, 84:5818–5822, 1987.

[122] T. Jukes and C. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, NY, 1969.

[123] L.A. Katz. Lateral gene transfers and the evolution of eukaryotes: Theories and data. *International Journal of Systematic and Evolutionary Microbiology*, 52:1893–1900, 2002.

[124] P.J. Keeling and W.M. Doolittle. Widespread and ancient distribution of a noncanonical genetic code in diplomonads. *Mol. Biol. Evol.*, 14:895–901, 1997.

[125] M.G. Kidwell. Lateral transfer in natural populations of eukaryotes. *Ann. Rev. Genet.*, 27:235–256, 1993.

[126] J. Kim and T. Warnow. Tutorial on phylogenetic tree estimation. In *Proc. 7th Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB99)*, 1999.

[127] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120, 1980.

[128] A.G. Kluge. Total evidence or taxonomic congruence: cladistics or consensus classification. *Cladistics*, 14:151–158, 1998.

[129] E.V. Koonin, K.S. Makarova, and L. Aravind. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annual Review of Microbiology*, 55:709–742, 2001.

[130] V. Krishnapillai. Horizontal gene transfer. *Journal of Genetics*, 75(2):219–232, 1996.

[131] V.N. Krylov. The role of horizontal gene transfer by bacteriophages in the origin of pathogenic bacteria. *Russian Journal of Genetics*, 39(5):483–504, 2003.

[132] V. Kunin and C.A. Ouzounis. The balance of driving forces during genome evolution in prokaryotes. *Genome Research*, 13(7):1589–1594, 2003.

[133] C.G. Kurland. Something for everyone—horizontal gene transfer in evolution. *Embo Reports*, 1(2):92–95, 2000.

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

[135] J.G. Lawrence. Gene transfer in bacteria: Speciation without species? *Theoretical Population Biology*, 61(4):449–460, 2002.

[136] J.G. Lawrence and H. Ochman. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.*, 44:383–397, 1997.

[137] J.G. Lawrence and H. Ochman. Reconciling the many faces of lateral gene transfer. *Trends in Microbiology*, 10(1):1–4, 2002.

[138] P. Legendre. Biological applications of reticulation analysis. *J. Classif.*, 17(2):191–195, 2000.

[139] P. Legendre, editor. *Special section on reticulate evolution*, volume 17 of *J. Classif.*, pages 153–195. Springer-Verlag, 2000.

[140] P. Legendre and V. Makarenkov. Reconstruction of biogeographic and evolutionary networks using reticulograms. *Syst. Biol.*, 51(2):199–216, 2002.

[141] D.A. Levin. 50 years of plant speciation. *Taxon*, 50(1):69–91, 2001.

[142] D.A. Levin, J. Francisco-Ortega, and R.K. Jansen. Hybridization and the extinction of rare plant species. *Conservation Biology*, 10(1):10–16, 1996.

[143] W.-H. Li. *Molecular Evolution*. Sinauer Assoc., 1997.

[144] D.A. Liberles, D.R. Schreiber, S. Govindarajan, S.G. Chamberlin, and S.A. Benner. The adaptive evolution database (taed). *Genome Biol.*, 2(8), 2001.

[145] C.R. Linder, B.M.E. Moret, L. Nakhleh, A. Padolina, J. Sun, A. Tholse, R. Timme, and T. Warnow. An error metric for phylogenetic networks. Technical Report TR-CS-2003-26, Univ. of New Mexico, 2003.

[146] L.E. Lindler, G.V. Plano, V. Burland, G.F. Mayhew, and F.R. Blattner. Complete dna sequence and detailed analysis of the Yersinia pestis kim5 plasmid encoding murine toxin and capsular antigen. *Infection and Immunity*, 66(12):5731–5742, 1998.

[147] B. Liu, J.M. Vega, and M. Feldman. Rapid genomic changes in newly synthesized amphiploids of Triticum and Aegilops. II. Changes in low-copy coding DNA sequences. *Genome*, 41(4):535–542, 1998.

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

[149] W. Maddison. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? *Evol.*, 44:304–314, 1990.

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

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

[152] V. Makarenkov and P. Legendre. From a phylogenetic tree to a reticulated network. *J. Comput. Biol.*, 2003. Accepted, to appear.

[153] V. Makarenkov, P. Legendre, and Y. Desdevises. Modelling phylogenetic relationships using reticulated networks. *Zoologica Scripta*, 2003. Accepted, to appear.

[154] K.S. Makarova, L. Aravind, M.Y. Galperin, N.V. Grishin, R.L. Tatusov, Y.I. Wolf, and E.V. Koonin. Comparative genomics of the archaea (euryarchaeota): Evolution of conserved protein families, the stable core, and the variable shell. *Genome Research*, 9(7):608–628, 1999.

[155] K.S. Makarova and E.V. Koonin. Comparative genomics of archaea: How much have we learned in six years, and what's next? *Genome Biology*, 4(8), 2003.

[156] M. Marron, K.M. Swenson, and B.M.E. Moret. Genomic distances under deletions and insertions. In *Proc. 9th Int'l Conf. Computing and Combinatorics (COCOON03)*, volume 2697 of *Lecture Notes in Computer Science*, pages 537–547, 2003.

[157] A.P. Martin. Choosing among alternative trees of multigene families. *Mol. Phyl. Evol.*, 16(3):430–439, 2000.

[158] E.P. Martins. Phylogenies and comparative data, a microevolutionary perspective. *Phil. Trans. R. Soc. Lond. B*, 349:85–91, 1995.

[159] S. Mathews and M.J. Donoghue. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science*, 286:947–950, 1999.

[160] L.A. McDade. Phylogenetic analysis when reticulate evolution is rampant—empirical evidence from neotropical Aphelandra (Acanthaceae). *American Journal of Botany*, 74(5):744–744, 1987.

[161] L.A. McDade. Hybrids and phylogenetic systematics II: The impact of hybrids on cladistic analysis. *Evol.*, 46:1329–1346, 1992.

[162] G.F. McGuire, F. Wright, and M.J. Prentice. A graphical method for detecting recombination in phylogenetic data sets. *Mol. Biol. Evol.*, 14:1125–1131, 1997.

[163] G.F. McGuire, F. Wright, and M.J. Prentice. A Bayesian model for detecting past recombination events and DNA multiple alignments. *J. Comput. Biol.*, 7(1-2):159–170, 2000.

[164] G.E. McKinnon, D.A. Steane, B.M. Potts, and R.E. Vaillancourt. Incongruence between chloroplast and species phylogenies in Eucalyptus subgenus monocalyptus (Myrtaceae). *American Journal of Botany*, 86(7):1038–1046, 1999.

[165] G. McVean, P. Awadalla, and P. Fearnhead. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160:1231–1241, 2002.

[166] C. Medigue, T. Rouxel, P. Vigier, A. Henaut, and A. Danchin. Evidence for horizontal gene transfer in E. coli speciation. *J. Mol. Biol.*, 222:851–856, 1991.

[167] T.J. Merritt and J.M. Quattro. Evidence for a period of directional selection following gene duplication in a neurally expressed locus of triosephosphate isomerase. *Genetics*, 159:689–697, 2001.

[168] T.H.M. Mes, R.M. Fritsch, S. Pollner, and K. Bachmann. Evolution of the chloroplast genome and polymorphic its regions in Allium subg. Melanocrommyum. *Genome*, 42(2):237–247, 1999.

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

[170] M.M. Miyamoto and W.M. Fitch. Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol.*, 44(1):64–76, 1995.

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

[172] B.M.E. Moret, U. Roshan, and T. Warnow. Sequence length requirements for phylogenetic methods. In *Proc. 2nd Int'l Workshop Algorithms in Bioinformatics (WABI02)*, volume 2452 of *Lecture Notes in Computer Science*, pages 343–356, 2002.

[173] I. Moszer, E.P.C. Rocha, and A. Danchin. Codon usage and lateral gene transfer in Bacillus subtilis. *Current Opinion in Microbiology*, 2(5):524–528, 1999.

[174] J.H. Nadeau and B.A. Taylor. Lengths of chromosome segments conserved since divergence of man and mouse. *Proc. Nat'l Acad. Sci., USA*, 81:814–818, 1984.

[175] L. Nakhleh, B.M.E. Moret, U. Roshan, K. St. John, J. Sun, and T. Warnow. The accuracy of phylogenetic methods for large datasets. In *Proc. 7th Pacific Symp. on Biocomputing (PSB02)*, volume 7, pages 211–222, 2002.

[176] L. Nakhleh, U. Roshan, K. St. John, J. Sun, and T. Warnow. Designing fast converging phylogenetic methods. *Bioinformatics*, 17(90001):S190–S198, 2001. ISMB01 Conference.

[177] L. Nakhleh, U. Roshan, K. St. John, J. Sun, and T. Warnow. The performance of phylogenetic methods on trees of bounded diameter. In O. Gascuel and B.M.E. Moret, editors, *Proc. 1st Int'l Workshop Algorithms in Bioinformatics (WABI01)*, volume 2149 of *Lecture Notes in Computer Science*, pages 214–226, 2001.

[178] L. Nakhleh, J. Sun, T. Warnow, R. Linder, B.M.E. Moret, and A. Tholse. Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. In *Proc. 8th Pacific Symp. on Biocomputing (PSB03)*, pages 315–326. World Scientific Pub., 2003.

[179] J.D. Nason, N.C. Ellstrand, and M.L. Arnold. Patterns of hybridization and introgression in populations of oaks, manzanitas and irises. *American Journal of Botany*, 79(1):101–111, 1992.

[180] C.L. Nesbo, Y. Boucher, and W.F. Doolittle. Defining the core of nontransferable prokaryotic genes: the euyarcheal core. *J. Mol. Evol.*, 53:340–350, 2001.

[181] B. Neuffer and P. Jahncke. Rapd analyses of hybridization events in cardamine (Brassicaceae). *Folia Geobotanica & Phytotaxonomica*, 32(1):57–67, 1997.

[182] H. Ochman. Lateral and oblique gene transfer. *Curr. Opin. Genet. Dev.*, 11(6):616–619, 2001.

[183] H. Ochman and I.B. Jones. Evolutionary dynamics of full genome content in Escherichia coli. *Embo J.*, 19(24):6637–6643, 2000.

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

[185] R.G. Olmstead and J.A. Sweere. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. *Syst. Biol.*, 43(4):467–481, 1994.

[186] S.P. Otto and J. Whitton. Polyploid incidence and evolution. *Annual Review of Genetics*, 24:401–437, 2000.

[187] R. Overbeek and N. Larsen *et al.* WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research*, 28(1):123–125, 2000.

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

[189] R. Page, editor. *Tangled Trees: Phylogeny, Cospeciation, and Coevolution*. U. Chicago Press, 2002.

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

[191] R. Page and M.A. Charleston. Reconciled trees and incongruent gene and species trees. In B. Mirkin, F.R. McMorris, F.S. Roberts, and A. Rzehtsky, editors, *Mathematical Hierarchies in Biology*, volume 37. American Math. Soc., 1997.

[192] R. Page and M.A. Charleston. Trees within trees: Phylogeny and historical associations. *Trends in Ecol. and Evol.*, 13:356–359, 1998.

[193] J.D. Palmer and W.F. Thompson. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell*, 29:537–550, 1982.

[194] P. Pamilo and M. Nei. Relationship between gene trees and species trees. *Mol. Biol. Evol.*, 5:568–583, 1998.

[195] C.A. Panetsos and H.G. Baker. The origin of variation in 'wild' Raphanus sativus (Cruciferae) in California. *Genetica*, 38:243–274, 1967.

[196] C. Patterson, D.M. Williams, and C.J. Humphries. Congruence between molecular and morphological phylogenies. *Annu. Rev. Ecol. Syst.*, 24:153–188, 1993.

[197] P.J. Planet. Reexamining microbial evolution through the lens of horizontal transfer. In R. DeSalle, G. Giribet, and W. Wheeler, editors, *Molecular Systematics and Evolution: Theory and Practice*, pages 247–270. Birkhauser Verlag, 2002.

[198] C.P. Ponting. Plagiarized bacterial genes in the human book of life. *Trends in Genetics*, 17(5):235–237, 2001.

[199] D. Posada and K.A. Crandall. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proc. Nat'l Acad. Sci., USA*, 98:13757–13762, 2001.

[200] D. Posada and K.A. Crandall. Intraspecific gene geneaologies: trees grafting into networks. *Trends in Ecol. and Evol.*, 16(1):37–45, 2001.

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

[202] D. Posada, K.A. Crandall, and E.C. Holmes. Recombination in evolutionary genomics. *Annu. Rev. Genet.*, 36:75–97, 2002.

[203] M.A. Ragan. Detection of lateral gene transfer among microbial genomes. *Current Opinion in Genetics & Development*, 11(6):620–626, 2001.

[204] M.A. Ragan. On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Letters*, 201:187–191, 2001.

[205] A. Rambaut and N. C. Grassly. Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comp. Appl. Biosci.*, 13:235–238, 1997.

[206] L.A. Raubeson and R.K. Jansen. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. *Science*, 255:1697–1699, 1992.

[207] J. Raymond and R.E. Blankenship. Horizontal gene transfer in eukaryotic algal evolution. *Proc. Nat'l Acad. Sci., USA*, 100(13):7419–7420, 2003.

[208] L.H. Rieseberg. Homoploid reticulate evolution in Helianthus (Asteraceae): evidence from ribosomal genes. *American J. Bot.*, 78(9):1218–1237, 1991.

[209] L.H. Rieseberg. The role of hybridization in evolution—old wine in new skins. *American Journal of Botany*, 82(7):944–953, 1995.

[210] L.H. Rieseberg. Distribution of spontaneous plant hybrids. *Proc. Nat'l Acad. Sci., USA*, 93:5090–5093, 1996.

[211] L.H. Rieseberg. Hybrid origins of plant species. *Annu. Rev. Ecol. Syst.*, 28:359–389, 1997.

[212] L.H. Rieseberg. Chromosomal rearrangements and speciation. *Trends in Ecol. and Evol.*, 16(7):351–358, 2001.

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

[214] L.H. Rieseberg and S.E. Carney. Plant hybridization. *New Phytologist*, 140(4):599–624, 1998.

[215] L.H. Rieseberg, R. Carter, and S. Zona. Molecular tests of the hypothesized hybrid origin of two diploid Helianthus species (Asteraceae). *Evol.*, 44:1498–1511, 1990.

[216] L.H. Rieseberg, H.C. Choi, R. Chan, and C. Spore. Genomic map of a diploid hybrid species. *Heredity*, 70:285–293, 1993.

[217] L.H. Rieseberg, H.C. Choi, and D. Ham. Differential cytoplasmic versus nuclear introgression in Helianthus. *J. Heredity*, 82:489–493, 1991.

[218] L.H. Rieseberg and N.C. Ellstrand. What can molecular and morphological markers tell us about plant hybridization? *Crit. Rev. Plant. Sci.*, 12(3):213–241, 1993.

[219] L.H. Rieseberg and C.R. Linder. Hybrid classification: Insights from genetic map-based studies of experimental hybrids. *Ecology*, 80:361–370, 1999.

[220] L.H. Rieseberg and R.D. Noyes. Genetic map-based studies of reticulate evolution in plants. *Trends in Plant Science*, 3(7):254–259, 1998.

[221] L.H. Rieseberg, B. Sinervo, C.R. Linder, M.C. Ungerer, and D.M. Arias. Role of gene interactions in hybrid speciation: Evidence from ancient and experimental hybrids. *Science*, 272(5262):741–745, 1996.

[222] L.H. Rieseberg, C. van Fossen, and A.M. Desrochers. Hybrid speciation accompanied by genomic reorganization in wild sunflowers. *Nature*, 375(6529):313–316, 1995.

[223] L.H. Rieseberg and D.A. Warner. Electrophoretic evidence for hybridization between Tragopogon mirus and T. miscellus (Compositae). *Syst. Biol.*, 12:281–285, 1987.

[224] L.H. Rieseberg and J.F. Wendel. Introgression and its consequences in plants. In R. Harrison, editor, *Hybrid Zones and the Evolutionary Process*, pages 70–109. Oxford Univ. Press, 1993.

[225] L.H. Rieseberg, J. Whitton, and C.R. Linder. Molecular marker incongruence in plant hybrid zones and phylogenetic trees. *Acta Botanica Neerlandica*, 45(3):243–262, 1996.

[226] K. Ritland and J.E. Eckenwalder. Polymorphism, hybridization, and variable evolutionary rate in molecular phylogenies. In D. E. Soltis, P. S. Soltis, and J. J. Doyle, editors, *Molecular Systematics of Plants*, pages 404–429. Chapman and Hall, New York, 1992.

[227] D.R. Robinson and L.R. Foulds. Comparison of phylogenetic trees. *Math. Biosciences*, 53:131–147, 1981.

[228] F.J. Rohlf. Phylogenetic models and reticulations. *J. Classif.*, 17(2):185–189, 2000.

[229] A. Rokas and P.W.H. Holland. Rare genomic changes as a tool for phylogenetics. *Trends in Ecol. and Evol.*, 15:454–459, 2000.

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

[231] U.L. Rosewich and H.C. Kistler. Role of horizontal gene transfer in the evolution of fungi. *Annual Review of Phytopathology*, 38:325–, 2000.

[232] U. Roshan, B.M.E. Moret, T. Warnow, and T. Williams. Performance of supertree methods on various dataset decompositions. In O.R.P. Bininda-Emonds, editor, *Phylogenetic Supertrees*. Kluwer Publications, 2003.

[233] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.

[234] S.L. Salzberg and J.A. Eisen. Lateral gene transfer or viral colonization? *Science*, 293:1048, 2001.

[235] S.L. Salzberg, O. White, J. Peterson, and J.A. Eisen. Microbial genes in the human genome—lateral transfer or gene loss? *Science*, 292(5523):1903–1906, 2001.

[236] T. Sang and Y. Zhong. Testing hybridization hypotheses based on incongruent gene trees. *Syst. Biol.*, 49(3):422–434, 2000.

[237] D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.*, 5:555–570, 1998.

[238] S. Sattath and A. Tversky. Additive similarity trees. *Psychometrika*, 42(3):319–345, 1977.

[239] M.H. Schierup and J. Hein. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156:879–891, 2000.

[240] E.E. Schilling and J.L. Panero. Phylogenetic reticulation in subtribe Helianthinae. *American Journal of Botany*, 83(7):939–948, 1996.

[241] K. Schwenk, A. Ender, and B. Streit. What can molecular markers tell us about the evolutionary history of daphnia species complexes. *Hydrobiologia*, 307(1–3):1–7, 1995.

[242] P.M. Sharp and W.H. Li. The codon adaptation index—a measure of directional synonymous codon usage bias and its potential applications. *Nucleic Acids Res.*, 15:1281–1295, 1987.

[243] P.M. Sharp, D.C. Shields, K.H. Wolfe, and W.H. Li. Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science*, 246:808–810, 1989.

[244] A.J. Shaw and B. Goffinet. Molecular evidence of reticulate evolution in the peatmosses (Sphagnum), including S. ehyalinum sp. nov. *Bryologist*, 103(2):357–374, 2000.

[245] A. M. Shedlock and M. Okada. SINE insertions: powerful tools for molecular systematics. *Bioessays*, 22:148–160, 2000.

[246] F.J. Silva, A. Latorre, and A. Moya. Why are the genomes of endosymbiotic bacteria so stable? *Trends in Genetics*, 19(4):176–180, 2003.

[247] J. Maynard Smith and N.H. Smith. Detecting recombination from gene trees. *Mol. Biol. Evol.*, 15:590–599, 1998.

[248] M. Smith. Analyzing the mosaic structure of genes. *J. Mol. Evol.*, 34:126–129, 1992.

[249] P.E. Smouse. Reticulation inside the species boundary. *J. Classif.*, 17(2):165–173, 2000.

[250] P.H.A. Sneath. Cladistic representation of reticulate evolution. *Syst. Zool.*, 24(3):360–368, 1975.

[251] P.H.A. Sneath. Reticulate evolution in bacteria and other organisms: How can we study it? *J. Classif.*, 17(2):159–163, 2000.

[252] Y.S. Song and J. Hein. Parsimonious reconstruction of sequence evolution and haplotype blocks: Finding the minimum number of recombination events. In *Proc. 3rd Int'l Workshop Algorithms in Bioinformatics (WABI03)*, volume 2812, pages 287–302. Springer-Verlag, 2003.

[253] M.S.M. Sosef. Hierarchical models, reticulate evolution and the inevitability of paraphyletic supraspecific taxa. *Taxon*, 46(1):75–85, 1997.

[254] R. Srikanth, J. Singh, and K.P. Raju. The chromospheric network—(a) network evolution viewed as a diffusion process. *Solar Physics*, 187(1):1–9, 1999.

[255] M.A. Steel. Recovering a tree from the leaf colourations it generates under a Markov model. *Appl. Math. Lett.*, 7:19–24, 1994.

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

[257] D.B. Stein, D.S. Conant, M.E. Ahearn, E.T. Jordan, S.A. Kirch, M. Hasebe, K. Iwatsuki, M.K. Tan, and J.A. Thomson. Structural rearrangements of the chloroplast genome provide an important phylogenetic link in ferns. *Proc. Nat'l Acad. Sci., USA*, 89:1856–1860, 1992.

[258] D.B. Stein, J.D. Palmer, and W.F. Thompson. Structural evolution and flip-flop recombination of chloroplast DNA in the fern genus Osmunda. *Curr. Genet.*, 10:835–841, 1986.

[259] K. Strimmer, K. Forslund, B. Holland, and V. Moulton. A novel exploratory method for visual recombination detection. *Genome Biol.*, 4(5), 2003. R33.

[260] K. Strimmer and V. Moulton. Likelihood analysis of phylogenetic networks using directed graphical models. *Mol. Biol. Evol.*, 17:875–881, 2000.

[261] K. Strimmer, C. Wiuf, and V. Moulton. Recombination analysis using directed graphical models. *Mol. Biol. Evol.*, 18(1):97–99, 2001.

[262] D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis. Phylogenetic inference. In D.M. Hillis, B.K. Mable, and C. Moritz, editors, *Molecular Systematics*, pages 407–514. Sinauer Assoc., Sunderland, Mass., 1996.

[263] M. Syvanen. On the occurrence of horizontal gene transfer among an arbitrarily chosen group of 26 genes. *J. Mol. Evol.*, 54(2):258–266, 2002.

[264] M. Syvanen and C.I. Kado. *Horizontal Gene Transfer*. Chapman & Hall, 2000.

[265] N. Takahata. Gene geneaology in three related populations: consistency probability between gene and population trees. *Genetics*, 122:957–966, 1989.

[266] S.A. Teichmann and G. Mitchison. Is there a phylogenetic signal in prokaryote proteins? *J. Mol. Evol.*, 49:98–107, 1999.

[267] M.J. Telford, E.A. Herniou, R.B. Russell, and D.T.J. Littlewood. Changes in mitochondrial genetic codes as phylogenetic characters: Two examples from the flatworms. *Proc. Nat'l Acad. Sci., USA*, 97:11359–11364, 2000.

[268] A.R. Templeton, K.A. Crandall, and C.F. Sing. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, 132:619–633, 1992.

[269] R.S. Thorpe. Reticulate evolution and cladism—tests for the direction of evolution. *Experientia*, 38(10):1242–1244, 1982.

[270] A.J. Tosi, J.C. Morales, and D.J. Melnick. Paternal, maternal, and biparental molecular markers provide unique windows onto the evolutionary history of macaque monkeys. *Evolution*, 57(6):1419–1435, 2003.

[271] M.C. Ungerer, S.J.E. Baird, J. Pan, and L.H. Rieseberg. Rapid hybrid speciation in wild sunflowers. *Proc. Nat'l Acad. Sci., USA*, 95(20):11757–11762, 1998.

[272] J.D. van Elsas, S. Turner, and M.J. Bailey. Horizontal gene transfer in the phytosphere. *New Phytologist*, 157(3):525–537, 2003.

[273] B. Venkatesh, Y. Ning, and S. Brenner. Late changes in spliceosomal introns define clades in vertebrate evolution. *Proc. Nat'l Acad. Sci., USA*, 96:10267–10271, 1999.

[274] J.C. Venter, M.D. Adams, E.W. Myers, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351., 2001.

[275] A. von Haeseler and G.A. Churchill. Network models for sequence evolution. *J. Mol. Evol.*, 37:77–85, 1993.

[276] J.F. Wendel, A. Schnabel, and T. Seelanan. An unusual ribosomal dna sequence from gossypium gossypioides reveals ancient, cryptic, intergenomic introgression. *Mol. Phyl. Evol.*, 4(3):298–313, 1995.

[277] J.J. Wiens. Combining data sets with different phylogenetic histories. *Syst. Biol.*, 47:568–581, 1998.

[278] T. Williams and B.M.E. Moret. An investigation of phylogenetic likelihood methods. In *Proc. 3rd IEEE Symp. on Bioinformatics and Bioengineering BIBE'03*, pages 79–86. IEEE Press, 2003.

[279] C. Wiuf, T. Christensen, and J. Hein. A simulation study of the reliability of recombination detection methods. *Mol. Biol. Evol.*, 18(10):1929–1939, 2001.

[280] C.R. Woese. The universal ancestor. *Proc. Nat'l Acad. Sci., USA*, 95(12):6854–6859, 1998.

[281] C.R. Woese. Interpreting the universal phylogenetic tree. *Proc. Nat'l Acad. Sci., USA*, 97(15):8392–8396, 2000.

[282] C.R. Woese. On the evolution of cells. *Proc. Nat'l Acad. Sci., USA*, 99(13):8742–8747, 2002.

[283] M. Worobey. A novel approach to detecting and measuring recombination: New insights into evolution in viruses, bacteria, and mitochondria. *Mol. Biol. Evol.*, 18(1):1425–1434, 2001.

[284] R. Wyatt, I.J. Odrzykoski, A. Stoneburner, H.W. Bass, and G.A. Galau. Allopolyploidy in bryophytes: Multiple origins of Plagiomnium medium. *Proc. Nat'l Acad. Sci., USA*, 85:5601–5604, 1988.

[285] S.Z. Xu. Phylogenetic analysis under reticulate evolution. *Mol. Biol. Evol.*, 17(6):897–907, 2000.

[286] H. Zhu, J.F. Klemic, S. Chang, P. Bertone, A. Casamayor, K.G. Klemic, D. Smith, M. Gerstein, M.A. Reed, and M. Snyder. Analysis of yeast protein kinases using protein chips. *Nature Genetics*, 26(3):283–289, 2000.