7 Tools to Construct and Study Big Trees: A Mathematical Perspective

M. Steel
Biomathematics Research Centre, University of Canterbury,
Christchurch, New Zealand

CONTENTS

ABSTRACT

This chapter describes some of the ways in which mathematical techniques provide insights and useful tools for reconstructing and analysing large trees and networks that will be required for species rich groups. Classically, 'mathematical biology' conjures up images of complex systems of differential equations; however, in phylogenetics quite different approaches are appropriate. Discrete mathematics, particularly algorithmic methods, graph theory and combinatorics, along with probability theory and statistics are the tools of choice. We describe how they can be used to address a range of topical issues relevant to constructing a tree of life. Why might large trees be useful? Does it even make sense to talk about a tree in the presence of reticulate evolution (and how does tree incongruence allow one to quantify the extent of reticulation?). How can one best combine trees on overlapping sets of taxa into supertrees or supernetworks? And how is biodiversity lost as species go extinct? At present most of these questions have only partial answers that are undergoing constant revision. Their full solution is a challenge for the future that will involve a close interplay between (at least) four disciplines: biology, mathematics, statistics and computer science.

7.1 TREES (AND NETWORKS) OF LIFE

7.1.1 Introduction and Terminology

It is forty years since the first phylogenetic trees of vertebrates were constructed from amino acid sequence data for cytochrome c[1]. Since then our understanding of life has undergone continual revision thanks to a stream of technical advances, the discovery of new data types (DNA sequences, SINEs, gene order, AFLPs, etc.), better computing resources, new methodology and the enthusiasm of many researchers. Today with whole genome sequences, and gene databases with many billions of base pairs, the pace set by the early pioneers seems likely to continue. In short, we are living in a golden age for molecular phylogenetics.

So what role can mathematics and its associated fields, statistics and computer science, play? First it can help design faster and more accurate algorithms for reconstructing, analysing and comparing phylogenetic trees and networks. Many of the problems biologists would like solved are computationally intractable on large data sets ('NP-hard' in computer science jargon). However this often suggests variations on the problem that can be solved exactly and quickly. Mathematical techniques can also help explain why existing methods can be misleading for data that evolves under certain processes, and provide techniques to correct for this[2–4]; or help answer more basic questions, such as how much sequence data is needed to accurately reconstruct a large tree, or some deep divergence within it[5]? Perhaps the most tantalising questions are those that involve determining whether different processes necessarily lead to different signals in the data (and thereby allowing the data to test between these models), or whether some processes are effectively indistinguishable from each other. A further use of mathematics is in formalising ideas so that the assumptions required are explicit and unambiguous; often this leads to insights into the limits of what is possible with any method of tree building (as in Steel et al.[6]).

In this chapter we first outline why large trees might (or might not) be needed. In particular we give some statistical arguments in support of large phylogenies as a tool to better understand evolutionary processes and detail. We then consider the more basic question of whether the notion of a 'tree of life' can be rigorously defended, particularly in the face of reticulate evolution. We present the first of two new results in this chapter, namely a simple argument showing there is a well defined notion of a tree of life, even though reticulation and other processes may also play an important role. We then review some of the recent work that has helped set out a mathematical foundation for representing and studying reticulate evolution, and we describe some of the approaches that have been developed for constructing both supertrees and supernetworks. In Section 7.3 we consider one of the uses of large trees, namely the quantification of phylogenetic diversity, and its loss as species go extinct. Here we describe our second new result, namely we derive equations that show that the concave relationship between phylogenetic diversity and taxon sampling that have been reported in the literature is generic, and not particular to certain trees or branch lengths. The chapter ends with some brief concluding comments regarding progress and challenges that lie ahead in mathematical phylogenetics. Before going further, we introduce some terminology that will be used throughout the chapter.

Definitions. Mostly we follow the notation used in Semple and Steel[7]. We let $\mathcal{T}$ denote a *phylogenetic X*–tree, that is, a tree whose leaves comprise the set X of taxa (generally species or populations) under study, and whose remaining vertices (nodes) are of degree at least 3 (the degree of a vertex is the number of edges (branches) that are incident with it). The vertices at the tips are called *leaves*. If all the non-leaf vertices in a tree have three incident edges the tree is said to be *fully resolved* (sometimes called 'binary', these are the trees without polytomies, and so are maximally informative). We also deal with *rooted trees* which have some vertex (often the midpoint of an edge) distinguished as a root vertex; if we direct all the edges of the tree away from the root (so they are consistent with a time direction if the root is the ancestral taxon) then we can talk about the *clusters* of the tree, which are the subsets of X that lie below the different vertices of the tree (it is a classical result that any rooted phylogenetic tree can be uniquely reconstructed from its set of clusters). Often the edges of the trees (rooted or unrooted) will have a *(branch) length*, corresponding perhaps to the expected amount of evolutionary change on that edge. For a rooted tree, if the sum of these branch lengths from the root to any leaf is the same (for each leaf) we say the branch lengths satisfy a *molecular clock*. Of course, for any tree (including ones constructed from non-molecular data), if the vertices all have (temporal) dates and the leaves represent contemporaneous taxa, then assigning each branch the difference between the dates of its two vertices also gives branch lengths that satisfy a molecular clock.

7.1.2 Why Big Trees?

A variety of impressive tree of life projects such as the National Science Foundation funded CIPRES initiative[8] have set forth the challenge of reconstructing phylogenies of unprecedented size and scope. To build and analyse trees on thousands of taxa demands clever computational tools, including new supertree methods (these combine existing phylogenies into larger parent phylogenies) and more slick techniques for quickly and accurately reconstructing trees from primary genetic data. A further issue is how to cope with processes that can 'mess up' the tree such as lineage sorting, horizontal gene transfer, recombination and the formation of hybrid taxa (Rϕnsted et al., *Chapter 9*).

Given the substantial effort (and funding) being expended in this task, it is timely to discuss why such large trees are needed in the first place, beyond the obvious challenge ('because it's there').

After all, many of the central evolutionary questions in systematic evolutionary biology have come down to the relationship between three or four taxa or groups; the human-chimp-gorilla debate of the 1990s is one of many such examples. Four other arguments against the wholesale reconstruction of large trees can be summarised by the following viewpoints:

- **The traditionalist.** If one regards life as organised according to a Linnean hierarchy (species, genus, family, order etc.) then one need only build (smallish) trees *within* the taxonomic level of interest.
- **The logician.** If one knows the (rooted) tree for every set of three taxa, then this uniquely specifies the entire tree, so we need only ever build trees on three taxa.
- **The pessimist.** Large trees are only useful if they are accurate, yet to build very large accurate trees surely needs huge data, since the number of possible trees grows so quickly (superexponentially with the number of taxa). Moreover, a large tree would be so complex and difficult to visualise that it would obscure rather than illuminate interesting biological relationships and processes.
- **Life is a mess.** Although locally much of life may be described by a tree, a global tree of life does not exist; it is rather a tangled network comprising both vertical (tree-like) and horizontal (reticulation, horizontal gene transfer) evolution. Attempts to reconstruct a large tree of life are therefore misplaced.

Each viewpoint conveys some truth, but none is compelling in itself. For example, the traditionalist imposes a hierarchy a priori, whilst a more objective approach would allow the data to reveal how well it fits this scheme; experience with genetic data has shown that numerous relationships have turned out to violate a traditional classification (for example, see Maley and Marshall[9] and the references therein). The logician is technically correct (it is a mathematical theorem that the collection of rooted trees on triples of species determines uniquely the global tree), but it ignores a statistical reality: namely, one simply cannot infer all 3-taxon trees accurately due to sampling effects, model violation and site saturation on long branches (more about this below). Similarly, the view concerning life as a tangled network championed by Ford Doolittle and colleagues[10,11] is no doubt correct up to a point; however, it may still make sense to talk about a tree of life, and we describe below two ways this can be done.

In answer to the pessimist, simulations[12,13] (see also Bininda-Emonds and Stamatakis, *Chapter 6*) and analytical results[14] have suggested that trees for large numbers of taxa can be reconstructed with reasonable accuracy from moderate data. Furthermore, the accuracy of inferring historical relationships for a given set of taxa is often improved by sampling more taxa[15]. Additional taxa can break up long edges that give rise to misleading signals due to 'long branch attraction', and addition of taxa can also help address another problem in molecular phylogenetics, namely site saturation. This latter property of sequences results from a combination of high rates of nucleotide substitution and long time scales. It leads to the character state of a taxon at a leaf of the tree being essentially independent of the states in the rest of the tree (and therefore inferring the placement of this leaf in the tree is unreliable). Site saturation is a problem for any tree reconstruction method, it is not particular to methods such as maximum parsimony. Mathematical formulae based on information theory can provide a useful way to quantify the effect of site saturation, and the extent to which it can be ameliorated by including sequences for additional taxa[15,16]. These mathematical bounds typically require n (the number of taxa) to grow exponentially with the amount of evolutionary change, suggesting that the resolution of some deep divergences in a tree may require looking at very many taxa.

Regarding the pessimist's second point, rather than obscuring biological processes, large trees may instead be essential for testing them for two reasons. First, insufficient or biased sampling can mislead inferences; second, large trees may be required to perform meaningful statistical tests. One field of study where both these factors are important involves the analysis of tree shape (see Davies and Barraclough, *Chapter 10*; Hodkinson et al., *Chapter 17*). A number of studies[17–19] have considered various measures of 'tree balance' as a way of testing between different models of speciation (such as the Yule-Harding process, or the uniform (PDA) model). However if taxon sampling is incomplete and highly skewed by the availability of sequences or the interests of the particular investigator, then this may be the main influence on tree shape. Even if these problems can be overcome, it is difficult to reject one model in favour of another with small numbers of taxa unless the trees are extremely unbalanced; one frequently needs 50 or 100 taxa to decide for a given tree which model generated the data[20]. As models become more refined and the questions more delicate, it is clear that much larger trees, involving many hundreds or perhaps thousands of taxa, will be needed to tease these processes apart.

Another area where large trees are invaluable is in the study of models of DNA site substitution. Whilst some small trees (with just a few leaves) can reject certain models (for example, if some taxa are highly GC rich and others highly AT rich, then one can reject a stationary reversible process); for more delicate studies larger trees are needed. This is apparent, for example, in attempting to distinguish between covarion drift models of sequence evolution and rates-across-sites models. The two models 'look' very similar, in the sense that they produce similar site patterns, but trees involving many taxa are needed to tell them apart[21]. As models become more refined, it will require even more taxa to test between them; large trees may also provide a way to estimate underlying parameters in more standard models of sequence evolution (Elchanan Mossel, personal communication).

Finally, large trees are convenient from a statistical perspective, since various limit theorems exist for certain probability distributions. For example, the parsimony score of a random character on any fixed large fully resolved tree is normally distributed[22], and the parsimony score of a character evolved under a standard Markov process at low rates on a large tree is Poisson distributed[23]. For small trees, these convenient, familiar distributions must be replaced by a tedious ad hoc analysis or by simulations.

7.1.3 Is There a Tree of Life?

The notion of a 'tree of life' is central to evolutionary biology, yet problems arise when one tries to precisely formalise the notion. One issue is the thorny and long-standing question of what constitutes a 'species' (there are myriad definitions; see for example Wheeler and Meier[24]); another is that evolution has involved reticulate processes such as horizontal gene transfer and the formation of hybrid species (for example in certain plant, insect and animal species[25]), gene fusion and endosymbiosis. Furthermore, a species is not a single entity, but rather a population of individuals, and under sexual reproduction recombination can further complicate a tree-like description of ancestry. These and other details throw into doubt the plausibility of constructing any well defined notion of a tree of life, as noted by several authors (for example, Bapteste[10]). Wayne Maddison[26] has also explored the related question of what really constitutes a 'phylogeny'.

In this section we describe a simple mathematical result that shows how an underlying tree of life always can be defined (and exists) even in the presence of these various complications. To explain this result, first recall that a *hierarchy* C on X is a collection of subsets of X, containing X, and satisfying the property

$$A, B \in C \Rightarrow A \cap B \in \{\phi, A, B\},$$

that is, the sets in C are *nested*; if they have one or more species in common then one set is a subset of the other. It is a classical result that a hierarchy on X forms a tree whose leaves are labelled with subsets of X that partition X.

One can define a tree of life and avoid problematic notions concerning the definition of species by working at the level of individual organisms. Furthermore, all one needs to assume about evolution for this definition is that each organism on earth (now or in the past, excluding the first organisms) had at least one parent who originated before that organism. We show now how this assumption alone allows one to define an underlying tree. This tree does not represent the detailed history of ancestry of individual organisms (after all, sexual reproduction is inherently reticulate and so is represented by a pedigree graph rather than a tree). Rather, we describe a coarser structure, based on subsets of extant organisms that form nested clusters (and hence a tree) according to a property of their ancestry.

Of course this definition of a tree of life should not be taken too literally; the purpose here is much more modest, namely to show that one can define such a tree even in the face of the many complications of evolutionary biology mentioned above. Also, although the tree we discuss can in principle be computed (and in polynomial time), it requires knowing some detailed information about ancestry, and is unlikely to be feasible, at least at present. Nevertheless it is interesting to speculate what the tree we describe here looks like, and the approach may provide some enticing questions and fruitful approaches for future work.

Let X be the set of *all* extant living taxa, that is, all living organisms currently on Earth. Note that we are not regarding here X as a set of species or populations, but of individuals. Let Ω denote the (large but finite) collection of all living organisms throughout the history of life on Earth, and for any real number $t > 0$, let Ω_t denote all organisms that were alive anytime up to t years ago. Thus $X = \Omega_0$. For $x \in \Omega$, let $t(x)$ denote the time when organism x first arose (i.e., was born), measured, say, in years. Thus:

$$t(x) = \max\{t \geq 0 : x \in \Omega_t\}.$$

We suppose that each organism that has ever existed arose from one or more parent organism(s) either:

- By haploid reproduction, or
- By diploid (sexual) reproduction, or
- By some higher-level process involving two or more parent organisms (for example a complex endosymbiosis event), or
- By being part of the initial population P_0 that constituted an origin of life.

Stated formally, for each organism $x \in \Omega - P_0$ there is a subset $p(x)$ of Ω (of size 1, 2 or possibly higher) and with the following property (for some fixed $\varepsilon > 0$):

$$\text{(P1)}\quad \text{For } x, y \in \Omega, \quad \text{if } y \in p(x) \quad \text{then} \quad t(y) > t(x) + \varepsilon,$$

which merely formalises a familiar fact: *parents originate before their offspring*.

We refer to the triple $\mathcal{L} = ((\Omega_t, t \geq 0), P_0, p)$ as a *history of life*; it is essentially a pedigree on a grand scale showing all organisms and their parents back through time to the origin of life. There are two ways to define a natural system of clusters from $\mathcal{L}$, and we will see below (Proposition 7.1.1) that they are actually equivalent and form a tree.

For $a \in X$, let $P_t(a)$ denote the set of organisms that lived up to t years ago and which have a as an descendant. Formally, $P_t(a)$ is the subset of Ω_t consisting of those x in Ω_t for which there is a sequence of organisms, $x = x_0, x_1, x_2, \ldots, x_k = a$ (for some k) and with $x_i \in p(x_{i+1})$ for each i. We say that a set of extant organisms $A \subseteq X$ is an $\mathcal{L}$-cluster if it satisfies the following property: there is some time t for which the following holds for all $a, a' \in A$ and all $x \in X - A$:

$$P_t(a) \cap P_t(a') \neq \phi \quad \text{and} \quad P_t(a) \cap P_t(x) = \phi. \tag{7.1}$$

In words, this property states there was some time t (measured, say, in years) for which any two organisms in A shared an ancestor that lived at most t years ago, and such that any organism that was an ancestor of both an organism in A and an organism not in A lived more than t years ago. Let $C_{\mathcal{L}}$ denote the set of $\mathcal{L}$-clusters of X.

The second way to define a collection of subsets of X uses distances. Define a 'distance' $d_{\mathcal{L}}$ on X as follows: let $d_{\mathcal{L}}(x, y)$ be the first time before the present when x and y shared an ancestral organism. Formally:

$$d_{\mathcal{L}}(x, y) := \min\{t \geq 0 : P_t(x) \cap P_t(y) \neq \phi\}.$$

Note that d is symmetric ($d_{\mathcal{L}}(x,y) = d_{\mathcal{L}}(y,x)$) and finite ($d_{\mathcal{L}}(x,y) < \infty$), though $d_{\mathcal{L}}$ may fail to satisfy the triangle inequality (that is, $d_{\mathcal{L}}(x, y)$ may be larger than $d_{\mathcal{L}}(x,z) + d_{\mathcal{L}}(z,y)$). Given any distance function d on X, the *Apresjan clusters* of d are those subsets A of X for which

$$\max\{d(a,a') : a, a' \in A\} < \min\{d(a,x) : a \in A, x \in X - A\}.$$

It is well known, and easily shown, that for any distance function d (even if it fails the triangle inequality) the set of associated Apresjan clusters forms a hierarchy (see for example Bryant and Berry[27]; Devauchelle et al.[28]).

Proposition 7.1.1 *For any life history, $\mathcal{L}$ satisfying (P1), the set $C_{\mathcal{L}}$ is precisely the set of Apresjan clusters of $d_{\mathcal{L}}$ and so forms a hierarchy (or equivalently a rooted tree). Furthermore, $C_{\mathcal{L}}$ can be reconstructed from $d_{\mathcal{L}}$ in $O(|X|^2)$ time.*

Proof. Suppose that A is a $\mathcal{L}$-cluster of X. Let t_A denote a value of t for which (7.1) holds for all $a, a' \in A$ and $x \in X - A$. Then for all $a, a' \in A, x \in X - A$, we have $\max\{d_{\mathcal{L}}(a,a') : a, a' \in A\} \leq t_A$ and, by (P1), $\min\{d_L(a,x) : a \in A, x \in X - A\} > t_A$ so that A is an Apresjan cluster of $d_{\mathcal{L}}$. Conversely, suppose that A is an Apresjan cluster of $d_{\mathcal{L}}$. Let $t = \max\{d_{\mathcal{L}}(a,a') : a, a' \in A$. Then t satisfies (7.1) for all $a, a' \in A$ and $x \in X - A$, and so A is an $\mathcal{L}$-cluster. The last part of Proposition 7.1.1, follows from Corollary 2.1 of Bryant and Berry[27] that provides an explicit algorithm to reconstruct the Apresjan clusters from any distance function.

Although it may be reassuring to know that a tree of life can still be defined in the presence of complications such as reticulation, such a tree will inevitably miss much of the detail and richness of evolutionary history, and may be largely unresolved in places. To describe a second type of tree that underlies evolution, though at a higher species level, we first need to talk about the sorts of networks that have been proposed to represent reticulate evolution.

7.1.4 Modelling Reticulate Evolution

To represent evolutionary reticulation an appropriate tool is a directed acyclic graph or DAG. As with a rooted tree, this graph has an orientation (time direction); the 'acyclic' condition simply ensures that we cannot follow a path forward in time and arrive back at the same vertex (time travel).

Until recently approaches for representing reticulate evolution by DAGs were somewhat ad hoc. For example, a biologist might construct a tree and then insert some reticulate branches, perhaps guided by a minimization principle, as in Legendre and Makarenkov[29]. In the last few years intensive work by mathematicians and computer scientists has provided a more sound basis for representing and analysing reticulate evolution[30–35]. Formally, a hybrid phylogeny is a connected, directed graph $\mathcal{H} = (V, A)$ consisting of a set V of vertices, and a set A of arcs linking pairs of vertices, and with no directed cycles. There is generally a single 'root' vertex ρ of in-degree 0, and the set of vertices of out-degree 0 is the set X of extant taxa being classified (the *in-degree* of a vertex v is the number of arcs that end at v; the *out-degree* of v is the number of arcs that start at v). Usually some other minor conditions are imposed in the definition to avoid trivialities.

Let $\mathcal{H} = (V, E)$ be a hybrid phylogeny on X with root vertex ρ (Figure 7.1). Let V_T be the set of vertices of $\mathcal{H}$ consisting of ρ together with those vertices that do not lie on any undirected cycle. For a vertex v of V, let $c(v)$ denote the set of species on X for which there is a directed path from v to x (that is, the extant species for which v is an ancestor).

The proof of the next result is given in Baroni et al.[36] (related results appear in Gusfield and Bansal[32] and Huson et al.[37]). It essentially (and informally) says that if we 'squash down' all the parts of a hybrid network that are involved in reticulation, we always end up with a tree. Or put another way, any hybrid phylogeny can be thought of as a tree, with certain vertices expanded out to reveal reticulation.

Proposition 7.1.2 *Let $\mathcal{H} = (V, E)$ be a hybrid phylogeny on X. Then the collection* $\{c(v) : v \in V_T\}$ *'orms a hierarchy on X, and so forms a tree.*

An interesting optimization question is to quantify the extent of reticulate evolution. A natural measure of how much reticulation occurs in a hybrid phylogeny $\mathcal{H}$ is

$$h(\mathcal{H}) = \sum_{v \in V - \rho} (d^-(v) - 1)$$

where $d^-(v)$ is the in-degree of vertex v. It is easily seen that $h(\mathcal{H}) = 0$ precisely if $\mathcal{H}$ is a tree. For the hybrid phylogeny $\mathcal{H}$ in Figure 7.1, $h(\mathcal{H}) = 2$.

Now, suppose we have a collection of phylogenetic trees constructed from different genes. Assuming each tree correctly represents the history of the corresponding gene, then any incompatibility between the trees must be due to other processes such as reticulate evolution or lineage sorting. One can then ask for the fewest reticulate events required to explain the incompatibility. This question can be phrased more precisely as follows: Given rooted phylogenetic X–trees $T_1, T_2, \ldots, T_k$ find a hybrid phylogeny $\mathcal{H}$ that minimises $h(\mathcal{H})$ and displays $T_1, \ldots, T_k$. For example, consider Figure 7.1. This hybrid phylogeny displays both of the trees on the right, and, and this is the minimum value possible.

It was recently shown that, even for two rooted binary trees, this optimisation problem is computationally intractable[38]; nevertheless there are useful mathematical theorems that allow for lower bounds on $h(\mathcal{H})$ to be established, and these are often strong enough to pin down its exact

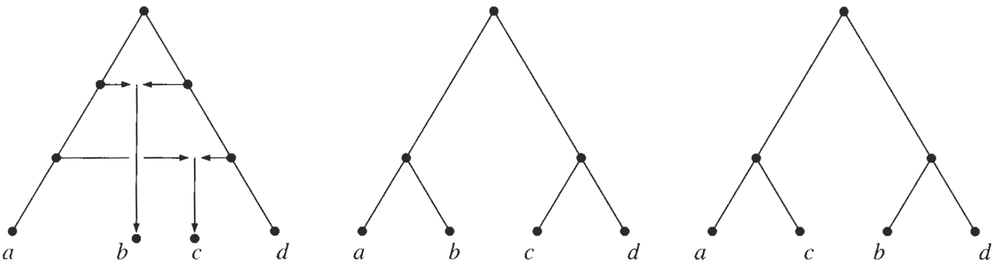

FIGURE 7.1 A hybrid phylogeny (left) that displays the two trees on the right.

value if the degree of reticulation is not too extreme[36,39]. A different, information-based approach to quantify reticulation, based on a notion of phylogenetic 'compression', has also been described recently by Ané and Sanderson[30].

7.2 CONSTRUCTING SUPERTREES AND SUPERNETWORKS

Large-scale trees and networks can be built from two types of input, either from existing phylogenetic trees (or networks), or directly from primary data, possibly combined from different sources. These two strategies have been referred to as a supertree (and supernetwork) versus a total evidence or supermatrix approach[40]. The latter viewpoint attracts considerable support from those who widely advocate maximum parsimony, although supertree approaches enjoy certain advantages, in particular, the availability of large databases of trees (such as TreeBASE[41]) to combine. Supertree approaches can also provide a useful 'divide-and-conquer' approach to tree reconstruction, where more complex models of sequence and genome evolution may preclude an analysis directly on large trees. In that case it may be more feasible to build many small trees, assigning them confidence values, and perhaps branch lengths, and then to combine these trees into a supertree or supernetwork. A further feature of supertree methods is that they provide a way of testing (and measuring) the extent to which the input trees may be incorrect, since if the taxa do indeed have a tree-like history, then the input trees should be consistent with some global tree. In practice, the nature of this inconsistency can be informative; for example, it may be due to one or two 'rogue' taxa that appear in different places in several input trees, or it may be that one tree has been built from poorly aligned sequence data.

Many methods for constructing supertrees have been developed recently. However, by far the most widely used supertree method is a traditional approach called matrix recoding with parsimony (MRP). This method recodes each tree as a set of partial binary characters, then combines them and applies standard maximum parsimony tree reconstruction to the resulting set of characters. The method's popularity is due more to the historical precedent and the availability of existing software than any compelling mathematical justification. However, it appears to produce reasonable trees and some basic desirable properties have been mathematically established. For example, if the input trees have the same leaf set, then any MRP tree will refine the strict consensus tree of the input trees (Theorem 2.16 of Bryant[42]).

Despite the widespread use of MRP, there has been some intensive development of alternative methods by computer scientists and mathematicians over the last six years or so. One of the problems with MRP is that it is computationally intensive; the point at which a large parsimony search is concluded is arbitrary, and there will generally be many (thousands) of output trees, so one typically then applies some consensus method to those trees to produce a single output. More direct algorithmic approaches to supertree methods include variations on the original (1981) BUILD algorithm of Aho et al.[43] including MinCutSupertree and its variants (see for example Bininda-Emonds[44]); these methods are provably fast (polynomial-time) and may be useful for building very large trees in realistic time.

7.2.1 Methods for Constructing Consensus Networks and Supernetworks

Methods to build phylogenetic networks (rather than trees) serve two related but distinct purposes, although the distinction has often become blurred in applications. These are:

- To exhibit conflicting signals in data, and uncertainty as to an underlying tree by providing support for alternative resolutions
- To explicitly model reticulate evolution due to processes such as the formation of hybrid species, hybridisation, recombination and so forth

Methods of the first type include Splits Graphs[45], NeighborNet[46], Median Neworks[47], Consensus Networks[48] and Z-closure networks[49]; all of which were developed by mathematicians. They provide useful representations of data. Methods of the second type include the supernetwork approach of Huson et al.[37] based on modifying split decomposition, and several other approaches[34,50,51]. A particularly simple and general approach to network construction is to construct the 'cover digraph' of a set of clusters (subsets of X); in some cases this can be an effective strategy for reconstructing a reticulate network when sufficient phylogenetic signals 'accumulate' in an evolutionary process[50]. A mathematically elegant technique to construct a network that displays a tree with multiple labels (arising from polyploidy) has also recently been developed[52].

7.2.2 Direct Methods for Analysing Genomic Data

A number of promising new techniques have also recently been developed for using genomic data to directly infer phylogeny and to better understand evolutionary processes. This area, often called phylogenomics, has caught the attention of many computer scientists. Methods range from those that deal just with the gene content of the taxa to more elaborate techniques for analysing differences in gene order (for recent surveys, see Moret et al.[53] and the comprehensive overview by Delsuc et al.[54]). Rare insertion events (such as SINEs) have also provided useful phylogenetic markers, and phylogenetic methods are currently being developed to use raw (non-aligned) genomic sequence based on relative information and compression measures (see for example Burstein et al.[55] and Otu and Sayood[56] and the references therein). Because of the potentially large state space (depending on how characters are coded) some genomic data, such as SINEs and gene order, often exhibit little or no homoplasy (parallel or convergent evolution), and so multistate maximum parsimony and related character-based methods are often quite efficient. The information content of such data is also generally high in the sense that the number of characters required to reconstruct a tree accurately can be shown mathematically to be relatively modest[5].

7.3 AN APPLICATION FOR LARGE TREES: PHYLOGENETIC DIVERSITY

A large tree, with branch lengths, provides a way to measure how much of the total evolutionary history is spanned by various subsets of the taxa. This measure, called 'phylogenetic diversity' (PD) and defined more precisely below, has been used as a comparative measure in biodiversity conservation, following its introduction by Dan Faith in 1992[57]. Subsequent authors[58–61] (see also the references therein) have explored its application further. In this section we describe some mathematical properties of this measure that are useful when applying it to large trees. In particular we describe how certain optimisation problems can be solved quickly by a greedy approach, and we also study the statistical properties of the process whereby PD is lost due to random extinction of taxa.

To define PD precisely, consider first an unrooted phylogenetic tree (one can think of such a tree as that obtained from any rooted phylogenetic tree by ignoring the root vertex; for precise definitions see Semple and Steel[7]). Let λ be an assignment of branch lengths to the edges of $\mathcal{T}$.

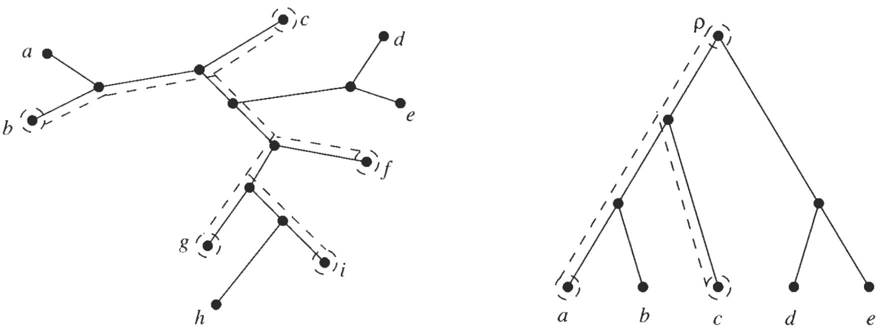

FIGURE 7.2 Left: For an unrooted tree $PD(W)$ for $W = \{b, c, f, g, i\}$ is the sum of the lengths of the dashed edges. Right: For a rooted tree $PD(W\ )$ for $W = \{b, c)$ is the sum of the lengths of the dashed edges.

Given a subset W of X, consider the induced phylogenetic W-tree, denoted $\mathcal{T}/W$ that connects just those species in W and its associated edge weighting λ_W which assigns to each edge e of $\mathcal{T}/W$ the sum of the $\lambda(e)$values over those edges of $\mathcal{T}$in the path that corresponds to e. The PD value of W, denoted $PD(W)$, is defined as

$$PD(W) := \sum_{e} \lambda_W(e)$$

where the summation is over all edges e in the tree $\mathcal{T}/W$. An example is illustrated in Figure 7.2 for $W = \{b,c,f,g,i\}$. Note that $PD(W)$ also depends on $(\mathcal{T}, \lambda)$, but we will think of these as fixed. Also, when $|W| = 1$ we set $PD(W) = 0$. In the case of a rooted phylogenetic tree, with root vertex ρ, we can regard the root as a leaf of an unrooted tree (with associated edge length 0), and then it is usual to define the phylogenetic diversity of a set W as $PD(W \cup \{\rho\})$ as illustrated in Figure 7.2 (this quantity for rooted trees has also been referred to as 'evolutionary history'[62]).

The PD score provides some indication of how much genetic variation each possible subset W contains in relation to the entire variation in the tree (by comparing $PD(W)$ to the total length of the tree $PD(X) = \sum_e \lambda(e)$). The PD score also turns out to have some interesting mathematical properties. In particular, it is possible to quickly find subsets of X of a given size that maximise PD by using a simple greedy approach. This was established for trees whose branch lengths satisfy a molecular clock[62] and extended to arbitrary trees[63]. The latter extension also allows for a subset Y of X of given size to be found that maximises $PD(Y) + a \cdot \sum_{y \in Y} f(y)$,where $f(y)$ is some value (or cost) of species y, and a is any (scale conversion) constant. In particular, one can ensure that certain species (including the root of a rooted tree) are always in the set Y (by giving them a large enough f value), and one can also find the taxa that are in all (or in none) of the maximal PD sets of given size. The fact that a fast (in this case greedy) approach works is vital for applications to large trees, since if one has a tree with (say) 1,000 taxa, and one wishes to find a subset of (say) 100 taxa that maximises the PD, then it is impossible for any computer to search all subsets of size 100 from the 1,000.

The combinatorial properties of PD have also been investigated[64], although for a different purpose, namely to show that the PD values of subsets of given size m suffice to uniquely determine the underlying tree (provided m is less than half the number of leaves of the tree). This approach has been developed further[65] to extend the popular neighbor joining tree reconstruction method so that it uses the PD values of taxa of given size (estimated, for example, by maximum likelihood) rather than just pairwise distance data.

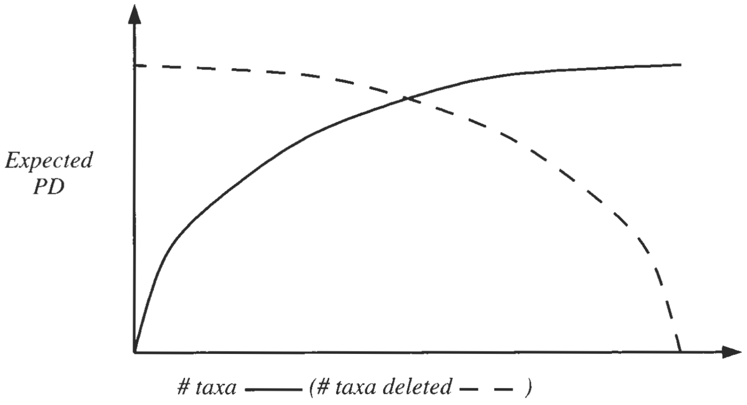

FIGURE 7.3 Concave relationships between PD gain/loss in a tree with addition/deletion of taxa.

We turn now to the statistical properties of PD as species go extinct. Nee and May[62] investigated the loss of PD as taxa are randomly deleted from random trees under a simple model in which each taxon is equally likely to be the next to go extinct (the 'field of bullets' model). The trees were generated by a random birth model with branch lengths that satisfy a molecular clock. They found a characteristic concave shape in the relationship between expected PD and the proportion of taxa deleted. This relationship was further investigated recently[66] on random deletion of taxa from certain biological trees. Once again the relationship between taxa deleted and PD was concave, as illustrated schematically in Figure 7.3. Recall that a sequence $x = (x_1, x_2, \ldots, x_n)$ of real numbers is *concave* if, when we let $\Delta x_r = x_r - x_{r-1}$, the following inequality holds for all r:

$$\Delta x_r - \Delta x_{r+1} \geq 0$$

and the sequence is strictly concave if the inequality is strict for all r; geometrically this means that the slope of the line joining adjacent points in the graph of x_r versus r is decreasing. Note that x_r is concave precisely if the complementary (reverse) sequence $y_r = x_{n-r}$ is. The significance of (strict) concavity for PD is that it says (informally) that most of the loss of PD comes near the end of an extinction process, as illustrated in Figure 7.3.

In this section we investigate the following question: is the concave relationship observed between the average PD and the number of taxa deleted particular to the trees (and the data or processes that generated them), or is it a generic property that applies to any tree with any set of branch lengths? We will see that the latter is true for any fully resolved tree with positive branch lengths. This makes intuitive sense because each interior branch survives until the point where there is no taxon that lies below that branch (which is likely to occur towards the end of a random extinction process). However, one could suspect that some trees with a certain assemblage of branch lengths might still lead to a violation of the concavity relationship, but the argument below rules this out. Perhaps the most satisfying aspect of the argument, however, is that we obtain exact expressions to describe the degree of concavity, in terms of the topology and branch lengths of the trees.

7.3.1 Concavity of Expected PD Is Generic

Consider a rooted phylogenetic tree having a leaf set X of size n. Let $E(\mathcal{T})$ denote the set of edges of $\mathcal{T}$, and let W be a random subset of taxa of size r sampled uniformly from X (for example, by selecting uniformly at random a set S of $n - r \geq 0$ elements of X and deleting them, in which case

$W = X - S$). For $r \in \{1, \ldots, n\}$ let $\mu_r = \mathbb{E}[PD(W)]$, the expected value of $PD(W)$ over all such choices of W. Equivalently,

$$\mu_r = \binom{n}{r}^{-1} \sum_{W \subseteq X:|W|=r} PD(W).$$

where $\binom{n}{r}$ is the binomial coefficient ($= \frac{n!}{r!(n-r)!}$), the number of ways of selecting r elements from a set of size n. Clearly $\mu_n = PD(X)$. For an edge e of $\mathcal{T}$ and a positive integer r let $\theta(e,r) = \frac{\binom{n-n_e}{r}}{\binom{n}{r}}$, where n_e denotes the number of leaves of $\mathcal{T}$ that lie 'below' (i.e., separated from the root by) e.

Proposition 7.3.1 *Consider a rooted phylogenetic tree $\mathcal{T}$ with an assignment λ of positive branch lengths. Then, for all $r \in \{0,\ldots,n\}$,*

$$\mu_r = PD(X) - \sum_{e \in E(\mathcal{T})} \lambda(e)\theta(e,r).$$

Proof. For each $e \in E(\mathcal{T})$, and W selected uniformly at random from all subsets of X of size r, consider the random variable $X_W(e)$ defined by setting

$$X_W(e) = \begin{cases} 1, & \text{if } W \text{ contains an element that lies below } e; \\ 0, & \text{otherwise.} \end{cases}$$

Then $PD(W) = \sum_{e \in E(\mathcal{T})} \lambda(e) X_W(e)$, and so

$$\mu_r = \mathbb{E}[PD(W)] = \sum_{e \in E(\mathcal{T})} \lambda(e)\mathbb{E}[X_W(e)]. \tag{7.2}$$

Now, $\mathbb{E}[X_W(e)] = 1 - \mathbb{P}[X_W(e) = 0]$, and the event $X_W(e) = 0$ occurs precisely if all the r elements of W are selected from amongst the leaves that are not below e. The probability of this occurring, when these r leaves are chosen randomly without replacement, is $\frac{\binom{n-n_e}{r}}{\binom{n}{r}}$, which is $\theta(e,r)$. Thus, $\mathbb{E}[X_W(e)] = 1 - \theta(e,r)$, which, combined with (7.2), establishes the Proposition.

To illustrate Proposition 7.3.1,

$$\mu_{n-1} = PD(X) - \frac{1}{n} \sum_{e \in E_{\text{ext}}(\mathcal{T})} \lambda(e),$$

where $E_{\text{ext}}(\mathcal{T})$ denotes the set of n (exterior) edges of $\mathcal{T}$ (leaves incident with a leaf).

For $r \in \{1, \ldots, n\}$, let $\Delta\mu_r = \mu_r - \mu_{r-1}$. Note that, since $\mu_0 = 0$, we have $\Delta\mu_1 = \mu_1$. For an edge e of $\mathcal{T}$, and $r \in \{1, \ldots, n-1\}$ let

$$\psi(e,r) := \frac{n_e(n_e - 1)}{r(r+1)} \cdot \frac{\binom{n-n_e}{r-1}}{\binom{n}{r+1}}.$$

We now describe the main consequence of Proposition 7.3.1. It shows that for any fully resolved tree PD decays in a strictly concave fashion as taxa are randomly deleted, and the only trees for which the decay of PD is linear are fully unresolved 'star' trees.

Corollary 7.3.2 *Consider a rooted phylogenetic tree $\mathcal{T}$ with an assignment λ of positive branch lengths. Then,*

1. For each $r \in \{1,\ldots,n-1\}$,

$$\Delta\mu_r - \Delta\mu_{r+1} = \sum_{e\in E(\mathcal{T})} \lambda(e)\psi(e,r).$$

In particular, μ is concave over this domain.

2. μ is strictly concave if and only if $\mathcal{T}$ has a cherry (i.e., there is an cluster of $\mathcal{T}$ that has precisely two leaves).
3. μ is linear if and only if $\mathcal{T}$ has no interior edges (i.e., is an unresolved 'star' tree).

Proof. From Proposition 7.3.1,

$$\Delta\mu_r - \Delta\mu_{r+1} = 2\mu_r - \mu_{r-1} - \mu_{r+1} = -\sum_{e\in E(\mathcal{T})} \lambda(e)[2\theta(e,r) - \theta(e,r-1) - \theta(e,r+1)]$$

and using a straightforward though tedious manipulation of (ratios of) binomial coefficients leads to the formula in the corollary.

For part (ii), if $\mathcal{T}$ has a cherry, let e be an edge with two leaves below it. Then $\psi(e,r) > 0$ for all $r \in \{1,\ldots,n-1\}$. Conversely, if $\Delta\mu_{n-1} - \Delta\mu_n > 0$, then there exists an edge e for which $\psi(e,n-1) > 0$, in which case $n_e = 2$, and so $\mathcal{T}$ has a cherry.

For part (iii), note that $\mathcal{T}$ is a star tree, if and only if $(n_e - 1) = 0$ for all edges e of $\mathcal{T}$, and this holds precisely if $\psi(e,r) = 0$ for all edges e of $\mathcal{T}$ and all values of r.

7.4 CONCLUDING COMMENTS

Mathematics has a long and successful history of application in the 'hard sciences', and Eugine Wigner[67] once talked of the "unreasonable effectiveness of mathematics" in physics. By contrast, the mathematician Gian Carlo-Rota wrote in 1986, "the lack of real contact between mathematics and biology is either a tragedy, a scandal or a challenge, it is hard to decide which"[68]. However, much has changed over the last two decades, and it seems that, in evolutionary biology, mathematics and related fields are starting to play a central role, and many of the techniques (such as NeighborNet, Split Decomposition, Median Networks and Hadamard transformation) have sprung from some elegant mathematical theory. In this chapter we have listed only some of these, and there are many others.

Despite these successes, many challenges lie ahead. For example, although many supertree methods have been proposed, very few also incorporate branch length information in the input trees and provide corresponding branch length estimates in the output tree (or network). Recent work by Stephen Willson[69,70] has provided a useful lead as to how this can be done, but much more work is needed. Similarly, supertree and supernetwork methods that take account of confidence estimates (or Bayesian posterior probabilities) on branches of the tree would seem to be desirable, and the Bayesian techniques described in Rϕnquist et al.[71] may point the way forward. A further challenge for the future will be the analysis of more complex models in molecular systematics, particularly since even comparatively simple 'mixture' processes of DNA site evolution can be problematic for standard computational approaches in statistics such as MCMC[72].

ACKNOWLEDGEMENTS

We thank the New Zealand Marsden Fund and the Allan Wilson Centre for Molecular Ecology and Evolution for supporting this research. I also thank Arne Mooers, Peter Lockhart, Klaas Hartmann, an anonymous referee and the editors for some helpful comments on an earlier version of this chapter.

REFERENCES

1. Fitch, W.M. and Margoliash, E., Construction of phylogenetic trees, *Science*, 155, 279, 1967.
2. Hendy, M.D. and Penny, D., A framework for the quantitative study of evolutionary trees, *Syst. Zool.,* 38, 297, 1989.
3. Lockhart, P.J. et al., Recovering evolutionary trees under a more realistic model of sequence evolution, *Mol. Biol. Evol.,* 11, 605, 1994.
4. Susko, E., Inagaki, Y., and Rogers, A.J., On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled, *Mol. Biol. Evol.* , 21, 1629, 2004.
5. Mossel, E. and Steel, M., How much can evolved characters tell us about the tree that generated them? in *Mathematics of Evolution and Phylogeny* , Gascuel, O., Ed., Oxford University Press, 2005, chap. 14.
6. Steel, M., Böcker, S., and Dress, A.W.M., Simple but fundamental limits for supertree and consensus tree methods, *Syst. Biol.,* 49, 363, 2000.
7. Semple, C. and Steel, M., *Phylogenetics,* Oxford University Press, 2003.
8. CIPRES, Building the Tree of Life: A national resource for phyloinformatics and computational phylogenetics (http://www.phylo.org).
9. Maley, L.E. and Marshall, C.R., The coming of age of molecular systematics, *Science*, 279(5350), 505, 1998.
10. Bapteste, E. et al., Do orthologous gene phylogenies really support tree-thinking? *BMC Evol. Biol.,* 5, 33, 2005.
11. Doolittle, W.F., Phylogenetic classification and the universal tree, *Science*, 284, 2124, 1999.
12. Salamin, N., Hodkinson, T.R., and Savolainen, V., Towards building the tree of life: a simulation study for all angiosperm genera, *Syst. Biol.,* 54, 183, 2005.
13. Zwickl, D.J. and Hillis, D.M., Increased taxon sampling greatly reduces phylogenetic error, *Syst. Biol.,* 51, 588, 2002.
14. Erdös, P.L. et al., A few logs suffice to build (almost) all trees (Part 1), *Rand. Struct. Algor.,* 14(2), 153, 1999.
15. Pollock, D.D. et al., Increased taxon sampling is advantageous for phylogenetic inference, *Syst. Biol.*, 51, 664, 2002.
16. Sober, E. and Steel, M., Testing the hypothesis of common ancestry, *J. Theor. Biol.,* 218, 395, 2002.
17. Aldous, D., Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, *Stat. Sci.,* 16, 23, 2001.
18. Chan, K.M.A. and Moore, B.R., Whole-tree methods for detecting differential diversification rates, *Syst. Biol.,* 51, 855, 2002.
19. Heard, S.B. and Mooers. A.O., The signatures of random and selective mass extinctions in phylogenetic tree balance, *Syst. Biol.,* 51, 889, 2002.
20. McKenzie A. and Steel, M., Distributions of cherries for two models of trees, *Math. Biosci.* 164, 81, 2000.
21. Lockhart, P.J. et al., How molecules evolve in Eubacteria, *Mol. Biol. Evol.,* 17, 835, 2000.
22. Steel, M.A., Goldstein, L., and Waterman, M., A central limit theorem for parsimony length of trees, *Adv. Appl. Prob.,* 28, 1051, 1996.
23. Steel, M., and Penny, D., Maximum parsimony and the phylogenetic information in multi-state characters, in *Parsimony, Phylogeny and Genomics* , Albert, V., Ed., Oxford University Press, 2005, chap. 9.
24. Wheeler, Q. and Meier, R., *Species Concepts and Phylogenetic Theory*, Columbia University Press, New York, 2000.
25. Mallet, J., Hybridization as an invasion of the genome, *Trends. Ecol. Evol.,* 20, 229, 2005.
26. Maddison, W., Gene trees in species trees, *Syst. Biol.,* 46, 523, 1997.

27. Bryant, D. and Berry, V., A structured family of clustering and tree construction methods, *Adv. Appl. Math.,* 27, 705, 2001.
28. Devauchelle, C., et al., Constructing hierarchical set systems, *Ann. Combin.,* 8, 441, 2004.
29. Legendre, P. and Makarenkov, V., Reconstruction of biogeographic and evolutionary networks using reticulograms, *Syst. Biol.,* 51, 199, 2002.
30. Ané, C. and Sanderson, M.J., Missing the forest for the trees: phylogenetic compression and its implications for inferring complex evolutionary histories, *Syst. Biol.*, 54(1), 146, 2005.
31. Baroni, M., Semple, C., and Steel. M., A framework for representing reticulate evolution, *Ann. Combin.,* 8, 391, 2004.
32. Gusfield, D. and Bansal, V., A fundamental decomposition theory for phylogenetic networks and incompatible characters, in *Proc. RECOMB 2005* , Miyato, S. et al. Eds., LNBI 3500, Springer-Verlag, Berlin Heidelberg, 2005, 217.
33. Huynh, T.N.D., Jansson, J., Nguyen, N.B. and Sung, W.-K., Constructing a smallest refining galled phylogenetic network, in *Proc. RECOMB 2005* , Miyato, S. et al. Eds., LNBI 3500, Springer-Verlag, Berlin Heidelberg, 2005, 265.
34. Moret, B. M. E. et al., Phylogenetic networks: modeling, reconstructibility, and accuracy, *IEEE/ACM Trans. Comput. Biol. Bioinf.,* 1, 1, 2004.
35. Song, Y. and Hein, J., On the minimum number of recombination events in the evolutionary history of DNA sequences, *J. Math. Biol.,* 48, 160, 2003.
36. Baroni, M., Semple, C. and Steel, M., Hybrids in real time, *Syst. Biol.,* 55, 46, 2006.
37. Huson, D.H. et al., Reconstruction of reticulate networks from gene trees, in *Proc. RECOMB 2005* , LNBI 3500 Miyano S. et al. Eds., Springer-Verlag, Berlin Heidelberg, 2005, 233.
38. Bordewich, M. and Semple, C., Computing the minimum number of hybridisation events for a consistent evolutionary history, Research Report (UCDMS2004/21), Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand, 2005.
39. Baroni, M. et al., Bounding the number of hybridisation events for a consistent evolutionary history, *J. Math. Biol.,* 51, 171, 2005.
40. Faith, D. P., From species to supertrees: Popperian corroboration and some current controversies in systematics, *Austr. Syst. Bot.,* 17, 1, 2004.
41. Sanderson, M.J. et al., TreeBASE: A prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life, *Am. J. Bot.* , 81, 183, 1994, (http://www.treebase.org/treebase).
42. Bryant, D., A classification of consensus methods for phylogenies, in *BioConsensus*, Janowitz, M., Lapointe, F.-J., McMorris, F.R., Mirkin, B., and Roberts, F.S. Eds., American Mathematical Society, 2003, 163.
43. Aho, A. V., Sagiv, Y., Szymanski, T. G., and Ullman, J. D., Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal on Computing* , 10, 405, 1981.
44. Bininda-Emonds, O.R.P., *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* , Kluwer Academic Publishers, Dordrecht, 2004.
45. Dress, A.W.M. and Huson, D.H., Constructing splits graphs, *IEEE/ACM Trans. Comput. Biol. and Bioinf.,* 1, 109, 2004.
46. Bryant, D. and Moulton, V., NeighborNet: an agglomerative algorithm for the construction of phylogenetic networks, *Mol. Biol. Evol.,* 21, 255, 2004.
47. Bandelt, H.-J., Forster, P., and Röhl, A., Median-joining networks for inferring intraspecific phylogenies, *Mol. Biol. Evol.* , 16, 37, 1999.
48. Holland, B. et al., Using consensus networks to visualize contradictory evidence for species phylogeny, *Mol. Biol. Evol.,* 21, 1459, 2004.
49. Huson, D. H. et al., Phylogenetic super-networks from partial trees, *IEEE/ACM Trans. Comput. Biol. Bioinf.,* 1, 151, 2004.
50. Baroni, M. and Steel, M., Accumulation phylogenies, *Ann. Combin.* , 10, 19, 2006.
51. Nakhleh, L., Warnow, T., and Linder, C.R., Reconstructing reticulate evolution in species—theory and practice. In *Proc. RECOMB 2004* , ACM, 2004, 337.
52. Huber, K.T. and Moulton, V., Phylogenetic networks from multi-labelled trees, *J. Math. Biol.* 52, 613, 2006.

53. Moret, B.M.E., Tang, J., and Warnow, T., Reconstructing phylogenies from gene-content and gene order data, in *Mathematics of Evolution and Phylogeny* Gascuel, O. Ed., Oxford University Press, chap. 12.
54. Delsuc, F., Brinkmann, H., and Philippe, H., Phylogenomics and the reconstruction of the tree of life, *Nature Rev. Genet.* , 6, 361, 2005.
55. Burstein, D. et al. Information theoretic approaches to whole genome phylogenies, in *Proc. RECOMB 2005*, Miyato, S. et al., Eds., LNBI 3500, Springer-Verlag, Berlin Heidelberg, 2005, 283.
56. Otu, H.H. and Sayood, K., A new sequence distance measure for phylogenetic tree construction, *Bioinf.*, 19, 2122, 2003.
57. Faith, D.P., Conservation evaluation and phylogenetic diversity, *Biol. Conserv.,* 61, 1, 1992.
58. Barker, G. M., Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation, *Biol. J. Linn. Soc.,* 76, 165, 2002.
59. Crozier, R.H., Dunnet, L.J., and Agapow, P.-M., Phylogenetic biodiversity assessment based on systematic nomenclature, *Evol. Bioinf. Online* , 1, 11, 2005.
60. Mooers, A.O., Heard, S.B., and Chrostowski, E., Evolutionary heritage as a metric for conservation, in *Phylogeny and Conservation* , Purvis, A., Brooks, T.L. and Gittleman, J.L. Eds., Cambridge University Press, Cambridge, 2005, 120.
61. Pavoine, S., Ollier, S., and Dufour, A.-B., Is the originality of species measurable? *Ecol. Lett.* , 8, 579, 2005.
62. Nee, S. and May, R.M., Extinction and the loss of evolutionary history, *Science*, 278, 692, 1997.
63. Steel, M., Phylogenetic diversity and the greedy algorithm, *Syst. Biol.,* 54, 527, 2005.
64. Pachter, L. and Speyer, D., Reconstructing trees from subtree weights,*Appl. Math. Lett.,* 17, 615, 2004.
65. Levy, D., Yoshida, R., and Pachter, L., Beyond pairwise distances: neighbor joining with phylogenetic diversity estimates, *Mol. Biol. Evol.* 23, 491, 2006.
66. Soutullo, A. et al., Distribution and correlates of Carnivore phylogenetic diversity across the Americas, *Animal Conserv.* , 8, 249, 2005.
67. Wigner, E.P., The unreasonable effectiveness of mathematics in the natural sciences, *Comm. Pure Appl. Math.,* 13, 1, 1960.
68. Kac, M., Rota, G.C., and Schwartz, J., *Discrete Thoughts*, Birkhauser, 1993.
69. Willson, S.J., Constructing rooted supertrees using distances, *Bull. Math. Biol.,* 66, 1755, 2004.
70. Willson, S.J., Unique solvability of certain hybrid networks from their distances, *Ann. Combin.,* 10, 165, 2005.
71. Ronquist, F., Huelsenbeck, J.P., and Britton, T., Bayesian supertrees, in *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* , Bininda-Emonds, O.R.P. Ed., Kluwer Academic Publishers, Dordrecht, 2004, 193.
72. Mossel, E. and Vigoda, E., Phylogenetic MCMC algorithms are misleading on mixtures of trees, *Science*, 309, 2207, 2005.