Tracing Evolutionary Links between Species

Mike Steel

Abstract. The idea that all life on earth traces back to a common beginning dates back at least to Charles Darwin's *Origin of Species*. Ever since, biologists have tried to piece together parts of this 'tree of life' based on what we can observe today: fossils and the evolutionary signal that is present in the genomes and phenotypes of different organisms. Mathematics has played a key role in helping transform genetic data into phylogenetic (evolutionary) trees and networks. Here, I will explain some of the central concepts and basic results in phylogenetics, which benefit from several branches of mathematics, including combinatorics, probability, and algebra.

1. WHAT IS PHYLOGENETICS? All living organisms on earth harbor within their DNA a signature of their evolutionary heritage. By studying patterns and differences between the genetic makeup of different species, molecular biologists are able to piece together parts of the story of how life today traces back to a common origin. In this way, many basic questions can be answered. When did animals and plants diverge? Are fungi more closely related to plants or animals? How and when did photosynthesis arise? What is the closest living animal to the whales? Does speciation occur in bursts or at a steady rate? Other topics are proving more difficult to resolve — for example, deciphering the earliest history of life on earth.

Similar questions arise for evolutionary processes in other fields such as epidemiology (e.g., the relationship between different strains of influenza or human immunodeficiency virus—HIV) and linguistics (e.g., how languages diverged from one another over time). In all these fields, the analysis relies on an underlying mathematical theory, grounded in combinatorics, algebra, and stochastic processes, with the concept of an evolutionary tree as a unifying object.

In this article, I describe a cross section of some of the key concepts in phylogenetics, which is the theory of reconstructing and analyzing trees, or more complex networks, from data observed at the present. I describe some combinatorial features of phylogenetic trees, namely their encoding by set systems, their enumeration, their generation under random models of evolution, and the way in which they can perfectly display discrete data. I then focus on tree reconstruction from data (discrete or distance based), which may not perfectly fit a tree. Such imperfect data can occur when data evolve along the branches of the tree under a random Markov model. I end by outlining how tree reconstruction is possible from this evolved data, but the choice of method requires care to avoid falling into a 'zone' of statistical inconsistency.

2. HIERARCHIES AND PHYLOGENETIC TREES. The 18th century Swedish taxonomist Carl Linneaus noticed that much of the living world can be nicely organized into a hierarchy in which groups of living organisms are either disjoint or nested [**30**]. For example, cats and dogs comprise disjoint classes of organisms, but both are subsets of the class of mammals. Formally, a *hierarchy* $\mathcal{H}$ on a finite set X is a collection of nonempty subsets of X with the property that any two elements of $\mathcal{H}$ are either nested (one is contained in the other) or disjoint. It will also be convenient here to

http://dx.doi.org/10.4169/amer.math.monthly.121.09.771
MSC: Primary 92D15

require that any hierarchy on X contains the set X and all its singleton subsets. Thus, $\mathcal{H}$ forms a hierarchy if it satisfies the two properties:

H1: for any two sets $A, B \in \mathcal{H}$ we have $A \cap B \in \{A, B, \emptyset\}$; and

H2: $\mathcal{H}$ contains the entire set X and each singleton set $\{x\}$ for all $x \in X$.

The second condition is harmless: If $\mathcal{H}$ is any collection of sets that satisfies **H1**, we can always add the extra elements mentioned by **H2** without violating **H1**.

To connect hierarchies with trees, recall first that a *tree* T is a connected graph (V, E) with no cycles. Often we will deal with rooted trees for which the edges are all directed away from some root vertex, and so each vertex has an in-degree and out-degree. We first define a *rooted phylogenetic X-tree* to be a tree T in which:

- X is the set of leaves (vertices of out-degree 0);
- all the arcs (directed edges) are directed away from some root vertex ρ;
- every nonleaf vertex has out-degree at least 2.

Figure 1(a) shows a simple biological example of a phylogenetic X-tree for a set X of five species; it reveals one relationship that is perhaps surprising to most non-biologists: Genetic data indicate that fungi are more closely related genealogically to animals than to plants. The interior vertices of a phylogenetic tree represent hypothetical ancestral species, with the root ρ being the "most recent common ancestor" of the species at the leaves.

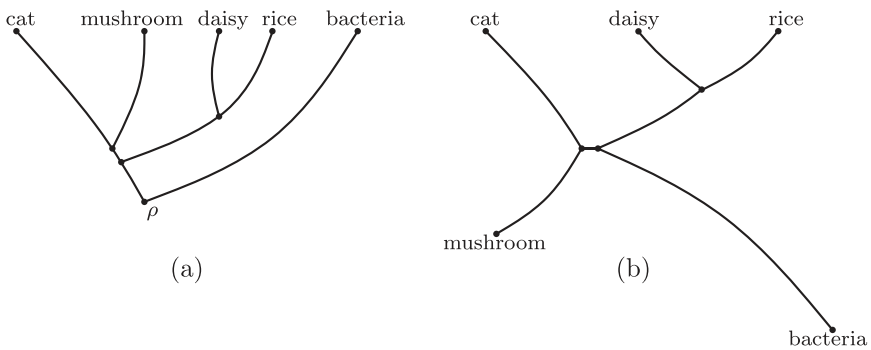

Figure 1. (a) A rooted phylogenetic X-tree, with root ρ. (b) The associated unrooted tree obtained by suppressing the root vertex.

We will think of two rooted X-trees as equivalent if they are isomorphic as rooted trees by an isomorphism that is the identity on X (i.e., trees are equivalent up to relabeling of the nonleaf vertices). Given a vertex v of T, the *cluster* associated with v is the subset of X that becomes separated from the root upon deletion of v. For the tree in Figure 1(a), the sets {cat, mushroom}, {daisy, rice} and {cat, mushroom, daisy, rice} are clusters.

Any collection $\mathcal{C}$ of subsets of X forms a directed graph, sometimes called the cover digraph of $\mathcal{C}$. The vertices of this graph are the elements of $\mathcal{C}$, and we place an arc from $B \in \mathcal{C}$ to $A \in \mathcal{C}$ precisely if B covers A (i.e., $A \subset B$ and there is no set $C \in \mathcal{C}$ with $A \subset C \subset B$). Now, the clusters of any phylogenetic X-tree T form a hierarchy on X, and the cover digraph of this hierarchy is isomorphic to T under the map that sends each vertex of T to the cluster associated with that vertex. Moreover, every hierarchy can be realized in this way, and nonequivalent phylogenetic trees give rise to different hiearchies. In other words, we have the following fundamental bijective correspondence between hierarchies and rooted phylogenetic X-trees (up to equivalence) [**16**].

Lemma 1. *A collection $\mathcal{C}$ of nonempty subsets of X is a hierarchy if and only if $\mathcal{C}$ is the set of clusters of some rooted phylogenetic X-tree T. Moreover, T is unique up to equivalence.*

The maximal hierarchies correspond to rooted phylogenetic X-trees in which every nonleaf vertex of T has out-degree 2, which are called *binary* trees (the tree in Figure 1(a) is an example). We will see shortly that such trees have exactly $2n - 1$ vertices where $n = |X|$, and so (by Lemma 1) this is the size of the largest hierarchy on a set of size n. Biologists often prefer trees to be binary since they show just one new lineage splitting off at a time; by contrast, a vertex of out-degree three or more represents what biologists call a polytomy (usually interpreted as uncertainty about the order of speciation events, rather than certainty about a sudden speciation event into multiple lineages).

The utility of viewing a rooted phylogenetic tree as a set system (a hierarchy) is illustrated by two questions biologists often face. Suppose we have a collection of different trees that estimate the evolutionary history of the same set of taxa. These trees might have been constructed by comparing genetic data across these species, but different choices of which genetic data to use (e.g., different genes) could have resulted in different tree estimates. In other words, while there might be one underlying and unknown true species tree that we wish to infer, the phylogenetic trees constructed from data will typically be merely imperfect estimates of this tree since the data evolve randomly, a topic we will discuss later. So two problems arise.

- How can we compare different phylogenetic X-trees?
- Can we combine different phylogenetic X-trees into some consensus tree?

The hierarchy link provides a very simple solution to both questions. First, observe that we can define a distance d between any two rooted phylogenetic X-trees T and T' by taking $d(T, T')$ to be the number of clusters that are present in one but not both of the trees T and T'. This distance d is called the Robinson–Foulds metric; it satisfies the triangle inequality and it can be computed quickly.

Turning to the consensus question, given a sequence of rooted phylogenetic X-trees $T_1, T_2, \ldots, T_k$, let $\mathcal{H}^{>1/2}$ be the set of clusters that are present in more than half of the corresponding hierarchies. In other words, if $\mathcal{H}_i$ is the hierarchy on X corresponding to T_i then $\mathcal{H}^{>1/2} = \{C \in \bigcup_{i=1}^{k} \mathcal{H}_i : |\{j : C \in \mathcal{H}_j\}| > k/2\}$. The following lemma shows that $\mathcal{H}^{>1/2}$ forms the set of clusters of a tree, the so-called majority rule consensus tree.

Lemma 2. *$\mathcal{H}^{>1/2}$ forms a hierarchy and so corresponds to a rooted phylogenetic X-tree.*

Proof. Suppose $C, C' \in \mathcal{H}^{>1/2}$. By the pigeonhole principle, there must be some hierarchy $\mathcal{H}_j$ that contains both C and C'. Consequently, C and C' are either disjoint or one is nested in the other. As this holds for all $C, C' \in \mathcal{H}^{>1/2}$, condition **H1** holds. Moreover, **H2** holds also since $\{x\}$ and X are elements of $\mathcal{H}_j$ for every j and every $x \in X$. Thus, $\mathcal{H}^{>1/2}$ forms a hierarchy and so (by Lemma 1), corresponds to a rooted phylogenetic X-tree that is unique up to equivalence. ■

The majority rule consensus tree has the nice combinatorial property that it comes as "close as possible," on average, to the input trees $T_1, T_2, \ldots, T_k$ under the Robinson–Foulds metric; more precisely, it is a median tree T that minimizes $\sum_{i=1}^{k} d(T, T_i)$ [**10**].

So, we can compare and combine phylogenetic trees. And once a biologist has a tree, it can help answer questions of interest, like "how long ago did two given species

have a common ancestor?" or "how did some given characteristic that varies between species (e.g., brain size) evolve?" But first we need a tree. A fundamental problem in phylogenetics is how to reconstruct— or infer—a tree from data present in the species today. Biologists also want to know how accurate such a reconstructed tree is likely to be. Before delving into tree reconstruction, it helps to understand the combinatorics of trees better. We start by considering the connection between rooted and unrooted trees, followed by their enumeration, and shape.

3. BASIC PROPERTIES OF PHYLOGENETIC TREES. Rooted trees appeal to biologists since they show evolution happening in time, from the past to the present. But it is often more convenient to consider unrooted trees. One reason is that most methods for building trees from data are generally able to do so only up to the placement of the root and so produce unrooted trees (figuring out where the root goes requires further work). Also, from a mathematical perspective, unrooted trees are often the more natural object to consider. The choice to work with either rooted or unrooted trees is somewhat analogous to the distinction in classical geometry between the affine and projective settings (respectively), where also one viewpoint may have advantages over the other, depending on the questions at hand.

Definition: An *(unrooted) phylogenetic X-tree* is a tree T with leaf set X and with every interior (i.e., nonleaf) vertex of degree at least three. If the degree of every nonleaf vertex is exactly three, we say that T is a *binary phylogenetic X-tree*.

Here is a first property of such trees, which will be useful in the next section.

Lemma 3. *Any unrooted binary phylogenetic tree T with n leaves has $2n - 3$ edges.*

Proof. A standard result in elementary graph theory states that a connected graph is a tree if and only if the number N of vertices exceeds the number E of edges by 1. So if our tree T has i interior vertices, we have $N = i + n$, and so

$$E = i + n - 1. \tag{1}$$

Also, for any graph, the handshake lemma tells us that the sum of the degrees of the vertices of any finite graph equals $2E$ since each edge is counted twice in this sum. Now, for our tree T, the sum of the degrees is $[1 + 1 + \cdots + 1(n \text{ times})] + [3 + 3 + \cdots + 3(i \text{ times})]$, and so

$$2E = n + 3i. \tag{2}$$

Combining Equations (1) and (2), we see that $i = n - 2$, and so $N = i + n = 2n - 2$, which implies that $E = N - 1 = 2n - 3$. This completes the proof of Lemma 3. ■

Notation: We will let $R(X)$ and $U(X)$ be the sets of rooted and unrooted phylogenetic X-trees (up to equivalence), and $RB(X)$ and $UB(X)$ will denote the sets of rooted and unrooted binary phylogenetic X-trees. Thus, when X has just four elements, $UB(X)$ consists of the three *quartet trees*, while $U(X)$ has one additional star tree that has a single nonleaf vertex of degree four. When $X = [n] = \{1, \ldots, n\}$, we will write $R(n)$, $U(n)$, $RB(n)$, and $UB(n)$ for $R(X)$, $U(X)$, $RB(X)$, and $UB(X)$, respectively.

Unrooting and counting trees. Counting trees has a long tradition in mathematics, with Cayley's n^{n-2} formula from 1889 for the total number of trees on n labelled vertices the most famous example. Counting binary phylogenetic trees turns out to be easier, and it has a history that dates back to even earlier mathematical work, contemporary with Darwin [**42**]. To explain this, we first describe a close connection between rooted and unrooted phylogenetic trees. There are two natural ways to associate an unrooted phylogenetic X-tree with a rooted tree.

Adding an outgroup: Take a rooted phylogenetic tree on $X - \{x\}$ and attach x to the root of T by a new edge. Species x is called an outgroup species.

Suppressing the root: Simply ignore the root vertex ρ; if it has degree 2 then delete it and identify its two incident edges; while if the root has degree at least three, then just treat this vertex as an interior vertex with no special root status. An example is shown in Figure 1.

Notice that the operation adding an outgroup provides a bijection

$$o : R(X - \{x\}) \to U(X)$$

that restricts to a bijection from $RB(X - \{x\})$ to $UB(X)$. On the other hand, suppressing the root results in a surjective map

$$s : R(X) \to U(X)$$

that restricts to a surjective map from $RB(X)$ to $UB(X)$. Moreover, the number of elements of $RB(X)$ that map to the same tree in $T \in UB(X)$ is the number of edges in T, which Lemma 3 tells us is $2n - 3$ (where $n = |X|$). These observations show us that

$$|RB(X)| = (2n - 3)|UB(X)| = (2n - 3)|RB(X - \{x\}|.$$

In particular, if $r(n)$ is the number of rooted binary phylogenetic trees on a leaf set of size n, then $r(n) = (2n - 3)r(n - 1)$, which, together with $r(2) = 1$, gives

$$r(n) = 1 \times 3 \times 5 \times \cdots \times (2n - 3).$$

This product of the odd numbers is often written as a double factorial (in this case, $(2n - 3)!!$). Notice that it can be expressed in terms of ordinary factorials and powers of 2 as follows:

$$r(n) = \frac{(2n - 2)!}{(n - 1)!2^{n-1}}. \tag{3}$$

Graph theorists may recognize this quantity: It is the number of perfect matchings of a complete graph on $2n - 2$ vertices. In other words, if there are $2n - 2$ people in a room, (3) counts the number of handshake scenarios in which each person shakes hands with precisely one other person. The bijection between this set of scenarios and set $RB(n)$ is an interesting but nontrivial exercise [**17**].

Applying Stirling's approximation $n! \sim \sqrt{2\pi} \cdot n^{n+\frac{1}{2}} e^{-n}$ to (3), reveals that $r(n)$ grows very rapidly. For example, $r(10)$ is around 34 million, while $r(30)$ is more than 10^{38}. Biologists often want to build trees for hundreds (or even thousands) of species; so it's no surprise that mathematics has an important role to play in this task, as it would be impossible to check each tree to see how well it might "fit the data."

There is another way to arrive at (3) by using generating functions. If we consider the formal power series $\varphi(x) = x + \sum_{n\geq 2} r(n)\frac{x^n}{n!}$ then

$$\varphi(x) = \frac{1}{2}\varphi(x)^2 + x$$

since deleting the root of a tree $T \in RB(n)$ for $n \geq 2$ results in an unordered pair of rooted binary trees (one or both of which might be an isolated leaf) with leaf sets that partition $[n]$ into two parts. Solving this quadratic equation gives $\varphi(x) = 1 - \sqrt{1-2x}$, from which $r(n)$ pops out as $n!$ times the coefficient of x^n in the Taylor expansion of $\sqrt{1-2x}$. While this is a more complicated derivation, generating functions turn out to be very useful in other applications—for example, in deriving exact explicit formulae for the number of forests of rooted binary trees on a given leaf set.

For the number $u(n)$ of unrooted binary trees on a leaf set of size n, the bijection o described above gives $u(n) = r(n-1) = (2n-5)!!$ Nonbinary phylogenetic trees (rooted and unrooted) can also be counted using recursions, but a closed-form expression like that for binary trees has proved elusive.

Tree shapes. If we ignore the labeling of the leaves of a rooted or unrooted phylogenetic tree, we obtain a tree shape. For example, when $n = 4$, there are two rooted binary tree shapes: the fork tree shape and the pectinate tree shape, shown in Figure 2(a, b). Biologists are interested in the shapes of trees since they shed light on the process of speciation and extinction in evolution.

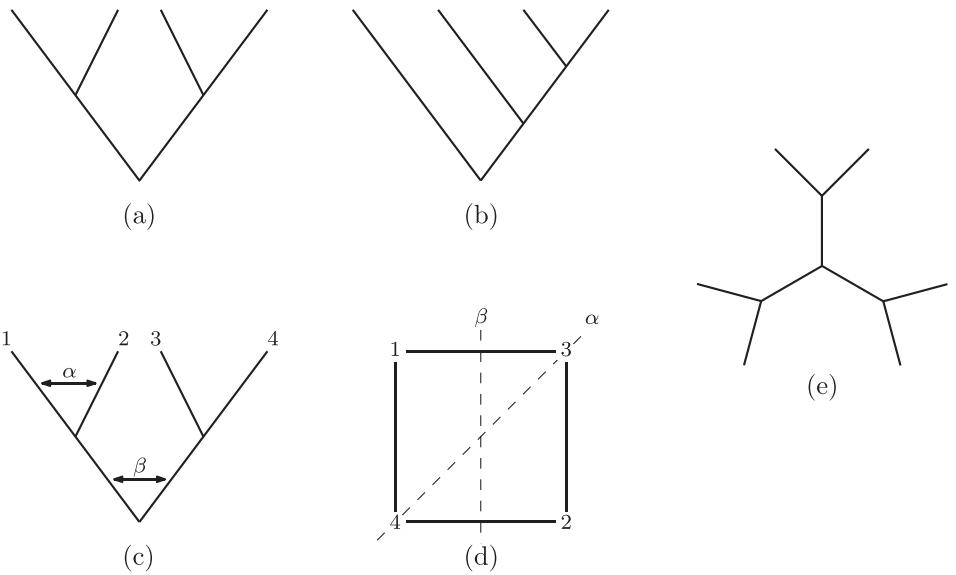

Figure 2. The two tree shapes for rooted binary trees on four leaves: (a) the fork and (b) the pectinate tree shape. The stabilizer subgroup of a phylogenetic tree having the fork shape (c) corresponds to the dihedral group of symmetries of a square (d). The two symmetries shown (α and β in (c)) correspond to reflections. In (e), an unrooted tree shape with a symmetry of order 3! about the central vertex is shown.

Elementary group theory provides a nice trick to count the number of phylogenetic X-trees of a given shape using the orbit-stabilizer theorem. Given a finite group G that acts on a set S, and an element $s \in S$, let $O(s) = \{g \cdot s : g \in G\} \subseteq S$ denote the orbit

of s under the action of G, and let $\mathrm{Stab}(s) = \{g \in G : g \cdot s = s\} \subseteq G$ be the stabilizer subgroup of G. Then the orbit-stabilizer theorem provides a bijection between the orbit of s and the cosets of $\mathrm{Stab}(s)$ in G, and so, in particular:

$$|O(s)| = \frac{|G|}{|\mathrm{Stab}(s)|}. \tag{4}$$

There is a natural action of the symmetric group Σ_n of permutations on $[n]$ on the set $R(n)$: Given $\sigma \in \Sigma_n$, simply permute the leaves of each tree T by replacing leaf x by leaf $\sigma(x)$. This action restricts to an action on the set $RB(n)$ of rooted binary trees, and so, by (4), the number of trees in $RB(n)$ that have the same shape as some tree T is $n!/|\mathrm{Stab}(T)|$. Now $\mathrm{Stab}(T)$ is a group of order $2^{s(T)}$ where $s(T)$ is the number of *symmetry vertices* of T—these are interior vertices for which the two subtrees of T that the vertex separates from the root have the same shape. For example, for a phylogenetic tree having the fork tree shape in Figure 2(a), T has three symmetry vertices and so $\mathrm{Stab}(T)$ is a group of order eight. This group is isomorphic to the dihedral group of rotational and reflectional symmetries of a square, as illustrated in Figure 2(c,d). In particular, for any set X of size four, there are precisely $4!/2^3 = 3$ rooted binary X-trees that have the shape of the fork tree; by contrast, the pectinate tree (Figure 2(b)) has only one symmetry vertex, and so there are 12 rooted binary phylogenetic X-trees of this shape.

For unrooted binary trees and nonbinary trees, similar formulae apply, though more complex symmetries arise; for example, an unrooted binary tree can have a two-fold symmetry about a central edge and, in the case of the tree shape shown in Figure 2(e), a symmetry of order 3! about a central vertex.

The shape of evolving trees: many roads lead to one distribution. As well as speciation, extinction has also played a major role in the history of life; after all, most species are extinct. Suppose we sample some subset X of species present today (species $a - e$ in Figure 3(i)) and then consider the minimal tree linking these species. This results in the so-called reconstructed tree illustrated in Figure 3(ii). Let's think of this as a rooted phylogenetic X-tree (ignoring the length of the edges). It turns out that, under very general assumptions concerning the speciation–extinction process, many models predict an identical and simple discrete probability distribution on $RB(X)$ [**29**].

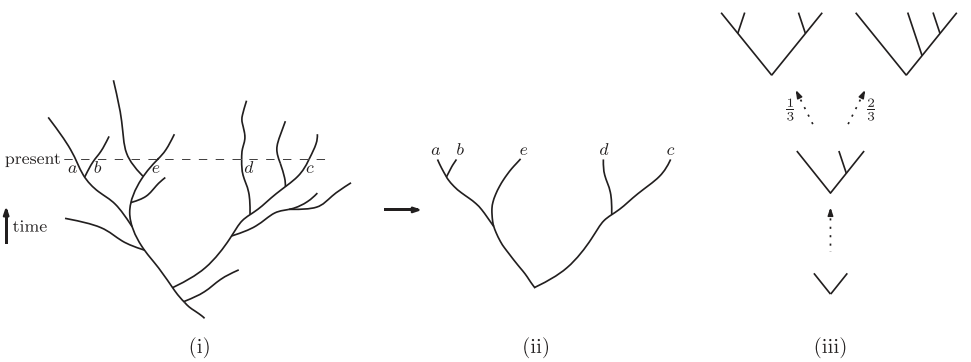

Figure 3. (i) A birth-death tree showing speciation and extinction. (ii) The associated "reconstructed tree," (iii) Growing a tree by the YH process.

This distribution is called the *Yule–Harding (YH)* distribution, and it is easily described as follows: To obtain a binary tree shape, we start with a tree shape on

two leaves and sequentially attach leaves—at each step attaching a new leaf to one of the leaf edges chosen uniformly at random from the tree constructed so far. For example, the probabilities of generating the fork and pectinate tree shapes are 1/3 and 2/3, respectively, since from the (unique) tree shape on three leaves, we can attach a new leaf to exactly one of the three leaf edges to obtain a fork tree shape or to either of two of these leaf edges to obtain a pectinate tree shape (see Figure 3(iii)).

Once we have built up a tree with n leaves in this way, we obtain a random tree shape on n leaves, and we can now label the leaves of this tree shape according to a permutation on $\{1, 2, \ldots, n\}$ chosen uniformly at random. This is the Yule–Harding probability distribution on $RB(n)$. Curiously, a quite different process that arises in population genetics and which proceeds backward in time (rather than forward, like Figure 3(iii)), also leads to the YH distribution when we ignore the length of the edges. This is the celebrated coalescent process of Sir John Kingman from the early 1970s.

We now explain how to compute the probability of a YH tree shape and of any rooted phylogenetic tree with this shape. First, let's grow a tree under the YH process until it has n leaves and then randomly select one of the two subtrees incident with the root (say the left-hand one since the orientation in the plane plays no role) and let Z_n denote the number of leaves in this tree. Remarkably, Z_n has a completely flat distribution.

Lemma 4. *Z_n has a uniform distribution between 1 and $n-1$, so*

$$\mathbb{P}(Z_n = i) = \frac{1}{n-1}, \quad \text{for } i = 1, \ldots, n-1.$$

Proof. The random process $Z_2, Z_3, \ldots$ can be exactly described as a special case of a classical process in probability called *Polya's Urn*. This consists of an urn that initially has a blue balls and b red balls. At each step, a ball is sampled uniformly at random and it is returned to the urn along with another ball of that same color. In our setting, $a = b = 1$ and blue corresponds to the left-hand subtree and red the right-hand subtree in the Yule–Harding tree. At each step, the uniform process of leaf attachment ensures that Z_n has exactly the same probability distribution as the number of blue balls in the urn after $n-2$ steps. It is well known, and easily shown by induction, that in Polya's Urn with $a = b = 1$, the proportion of blue balls has a uniform distribution. ■

This lemma provides the key to computing the YH probability of a tree exactly, as follows.

Proposition 1. *For any particular tree $T \in RB(n)$, the probability $\mathbb{P}_{YH}(T)$ of generating T under the YH model is given by*

$$\mathbb{P}_{YH}(T) = \frac{2^{n-1}}{n! \prod_{v \in I(T)} \lambda_v}$$

where $I(T)$ is the set of interior vertices of T and where λ_v is -1 plus the number of leaves of T that are descendants of v.

For example, for the tree in Figure 1(a), $\mathbb{P}_{YH}(T) = \frac{2^4}{5! \times 4 \times 3 \times 1^2} = \frac{1}{90}$, while for Figure 3(ii), $\mathbb{P}_{YH}(T) = \frac{1}{60}$.

Proof. Suppose that the two maximal subtrees T_1 and T_2 of T are of size k and $n-k$, where we may assume that $2k \leq n$. By Lemma 4, the probability of such a size distribution is $2/(n-1)$ if $2k < n$ and $1/(n-1)$ if $2k = n$. Conditional on this division, the number of ways to select leaf sets for T_1 and T_2 that partition $[n]$ is $\binom{n}{k}$ when $2k < n$, and $\frac{1}{2}\binom{n}{k}$ when $2k = n$ (the factor of $\frac{1}{2}$ recognizes that the order of T_1 and T_2 is interchangeable in T when they have the same number of leaves). By the Markovian nature of the YH process, each of these two subtrees also follows the YH distribution (in the special case $k = 1$ where T_1 is a single leaf, formally set $\mathbb{P}_{YH}(T_1) = 1$). This leads to the recursion

$$\mathbb{P}_{YH}(T) = \frac{2}{n-1}\binom{n}{k}^{-1}\mathbb{P}_{YH}(T_1)\mathbb{P}_{YH}(T_2),$$

from which Proposition 1 now follows by induction. ■

Notice that the YH process leads to a different probability distribution on $RB(n)$ from that obtained by simply selecting a tree uniformly at random from $RB(n)$, which would assign each $T \in RB(n)$ the probability $1/(2n-3)!!$ For example, the probability of obtaining a tree with the fork in Figure 2(a) has probability $\frac{1}{5}$ under a uniform distribution on $RB(4)$ (since only three of the 15 trees in $RB(4)$ have that shape) and $\frac{1}{3}$ under YH. Larger trees lead to more pronounced differences between the two distributions. For instance, in YH trees on n leaves, the expected number of edges between the root and a randomly selected leaf grows at the rate $\log(n)$, while for uniform binary trees, it grows at the rate $\sqrt{n}$. YH trees also tend to be more balanced than uniform trees, where balance refers to the average difference between the sizes of the two daughter subtrees in the tree, as one ranges over the interior vertices of the tree. For example, Proposition 1 shows that the probability that a tree with n leaves generated under the YH distribution has a single leaf adjacent to the root is $\frac{2}{n-1}$, while for the uniform distribution the corresponding probability is $\binom{n}{1}\frac{r(n-1)}{r(n)} = \frac{n}{2n-3}$, which converges to $\frac{1}{2}$ as n grows. It turns out that many real phylogenetic trees tend to have a degree of balance somewhere between that predicted by the YH and uniform distributions; explaining why has required some novel mathematical and statistical insights [**1, 29**].

4. TREES, SPLITS, AND CHARACTER DATA. In Section 3 we described a one-to-one correspondence between unrooted phylogenetic X-trees and rooted phylogenetic X-trees on $X - \{x\}$ (for any $x \in X$) and, thereby, to hierarchies on $X - \{x\}$. However, the choice of a particular element $x \in X$ is completely arbitrary, so we seek a more satisfactory way to describe an unrooted phylogenetic X-tree. This is based on the notion of an *X-split*, which is a bipartition of X into two nonempty parts (A and B, say), and written as $A|B$. Such a notion has clear biological meaning—for example, we can divide all life into the vertebrates and the invertebrates. Given any phylogenetic X-tree T, if we delete any particular edge e of T and consider the leaf sets of the two connected components of the resulting disconnected graph, we obtain a corresponding X-split, which we will refer to as a *split of T* that corresponds to e. For example, for each $x \in X$, every phylogenetic X-tree has the *trivial split* $\{x\}|X - \{x\}$, corresponding to the edge incident with leaf x.

Notice that any two splits $A|A'$ and $B|B'$ of the same phylogenetic X-tree have the property that one of the four intersections $A \cap B$, $A \cap B'$, $A' \cap B$, $A' \cap B'$ is empty. For example, the tree in Figure 1(b) has the splits $A|B = \{\text{cat, mushroom}\}|\{\text{daisy, rice, bacteria}\}$ and $A'|B' = \{\text{cat, mushroom, bacteria}\}|\{\text{daisy, rice}\}$ so in this case,

$A \cap B' = \emptyset$. If a collection Σ of X-splits has this property, we say that Σ is *pairwise compatible*. This is the unrooted analogue of the hierarchy property **H1**, so it is not surprising that Lemma 1 has an equivalent formulation for unrooted trees.

> A collection Σ of X-splits is the set of splits of some unrooted phylogenetic X-tree T if and only if Σ is pairwise compatible and contains the trivial splits. Moreover, T is uniquely determined up to equivalence by Σ.

We can extend the notion of splits further. Instead of deleting a single edge, we may delete any nonempty set E' of edges from a tree, and consider the resulting $|E'| + 1$ components of the disconnected graph $T - E'$. This gives an equivalence relation $\sim$ on X where $x \sim y$ precisely if x and y are connected in $T - E'$. The equivalence classes of $\sim$ comprise a partition of X into at most $|E'| + 1$ parts. Any such partition of X that can be obtained in this way is said to be *convex* on T, a concept that is relevant to the next part of the story.

Characters, homoplasy, and a perfect phylogeny. A function from the set of species X into some set S of r states is referred to by biologists as an r-state *character* on X. For example, $f(x)$ might be a morphological character that describes the number of legs that species x has or a genetic character the describes the nucleotide at a particular position in a genetic sequence for species x. That is, $f(x)$ describes some characteristic of x that we compare across other species in X. A hypothetical example of four characters across a set of eight well-known species is provided in the table below and will serve to illustrate several ideas that follow. If we regard each of the possible states as (say) letters of the alphabet, then we can associate a four-letter word to each species.

Table 1.

Species	Character	1	2	3	4
Kangaroo		T	R	U	E
Chimpanzee		B	R	E	T
Human		B	R	O	E
Gorilla		C	O	E	E
Hippopotamus		C	A	P	O
Whale		C	A	U	P
Lion		D	R	A	O
Tiger		D	R	U	G

If a phylogenetic X-tree describes the evolution of a set of species, a character tells us the states of the species at the leaves but not of the hypothetical ancestral species that correspond to the interior vertices of the tree. There are myriad ways to explain how the character could have evolved in the tree from some ancestral state at the root. It is possible that in a path from the root to a leaf a *reversal* occurs, where a state s_1 changes to state s_2 and later back to s_1; for example, in birds, wings first evolved and then in some species (e.g., kiwi) disappeared again. It is also possible for convergent evolution to occur, where state s_1 at some vertex changes to s_2 down two edge-disjoint paths that start from that vertex. Again, wings provide an example: From an ancestor of birds and mammals, wings evolved both in birds and in mammalian bats. A character whose evolution on a given tree can be explained without postulating any reversal or convergent events is said to be "homoplasy-free." Homoplasy-free evolution might be expected to hold when the number of potential states is very large, so each change is likely to be to a new state (for example, the order of genes on a chromosome under random rearrangement operations) or for certain genomic insertion data, such as

"Short Interspersed Nuclear Elements" (SINEs) in mammalian genetics (which helped establish that hippopotamus is the closest living species to whale [**34**]).

The notion of homoplasy-free can be defined more easily if we suppress the rooting of the tree and so consider unrooted trees. Formally, we will say that a character f on X is *homoplasy-free* on an unrooted phylogenetic X-tree T if $f : X \to S$ has an extension $F : V \to S$ to the set V of all vertices of T so that F is constant on the path between any two vertices with the same F-value. This is equivalent to requiring that for each $\alpha \in f(X)$, the subgraph of T induced by the set of vertices v with $F(v) = \alpha$ is connected. There are two other ways to characterize when a character $f : X \to S$ is homoplasy-free on a phylogenetic X-tree T:

- f has an an extension F to all the vertices of T for which F assigns different states to the endpoints of $|f(X)| - 1$ edges (equivalently, at most $|f(X)| - 1$ edges);
- the partition of X induced by the equivalence relation "$x \approx x'$ if and only if $f(x) = f(x')$" is convex on T (as defined just prior to Section 4).

Notice that homoplasy-free is considerably weaker than requiring that the actual evolution of the character on some rooting of the tree involved no reversals or convergent evolution—it merely requires that the character *could* have evolved in this way.

A sequence $(f_1, f_2, \ldots, f_k)$ of characters on X is said to have a *perfect phylogeny* if and only if there exists a phylogenetic X-tree on which each character is homoplasy-free (the tree is said to be a perfect phylogeny for those characters). We will see shortly that our eight-species example above forms a perfect phylogeny.

The computational problem of determining whether or not a collection of characters has a perfect phylogeny is NP-complete in general, but a polynomial-time algorithm exists when a bound is placed on either the number of characters or the number of states per character (r). In the special cases where $r = 2$ and $r = 3$, a collection of r-state characters has a perfect phylogeny if and only if every subset of size r of the characters has one. However, the if direction fails for larger values of r, as there is a set of $\lfloor \frac{r}{2} \rfloor \cdot \lceil \frac{r}{2} \rceil + 1$ characters on $r \geq 4$ states that do not have a perfect phylogeny even though every proper subset does [**45**]. The existence of a perfect phylogeny for a sequence of characters on X also has an attractive graph theoretic characterization involving chordal intersection graphs (for details, see [**44**]; more recent graph-based analysis of related approaches appears in [**6**]).

When a sequence of characters has a perfect phylogeny T, we can also ask when it is unique. A necessary condition for this is that T is binary; otherwise, we could arbitrarily resolve any vertex of T of degree greater than three and obtain a different tree on which all the characters were homoplasy-free. An interesting question now arises: What is the smallest number $h(n)$ so that for each $T \in UB(n)$ there is a sequence of $h(n)$ characters on $[n]$ that has T as a unique perfect phylogeny? If we restrict ourselves to binary characters, then

$$h(n) = n - 3$$

since for any $T \in UB(n)$, the characters that are homoplasy-free on T correspond precisely to the splits of T, and T is the unique perfect phylogeny for a sequence of such characters provided that all $n - 3$ nontrivial splits of T are represented (if one was missing we could contract the corresponding edge and still obtain a tree on which the characters were homoplasy-free). But what if we do not insist on restricting ourselves to two-state characters or r-state characters for any fixed r. Is it possible that $h(n)$ might grow more slowly than linearly with n; perhaps $\sqrt{n}$ or even $\log(n)$ characters might suffice? Surprisingly, it turns out that $h(n)$ is never more than four.

Theorem 1 (Four characters suffice). *For any binary phylogenetic X-tree T, there is a set S_T of at most four characters for which T is the* only *perfect phylogeny.*

An example of the set S_T from Theorem 1 is provided by the four hypothetical characters described for the eight species in Table 1. It is easily seen that the tree T shown in Figure 4 is a perfect phylogeny for this data set (this tree, incidentally, is the one biologists generally accept, rooted somewhere on the bottom edge). What is less obvious is that T is the only such perfect phylogeny for these four characters; moreover, the states at the interior vertices (shown in brackets) are uniquely determined by the homoplasy-free condition.

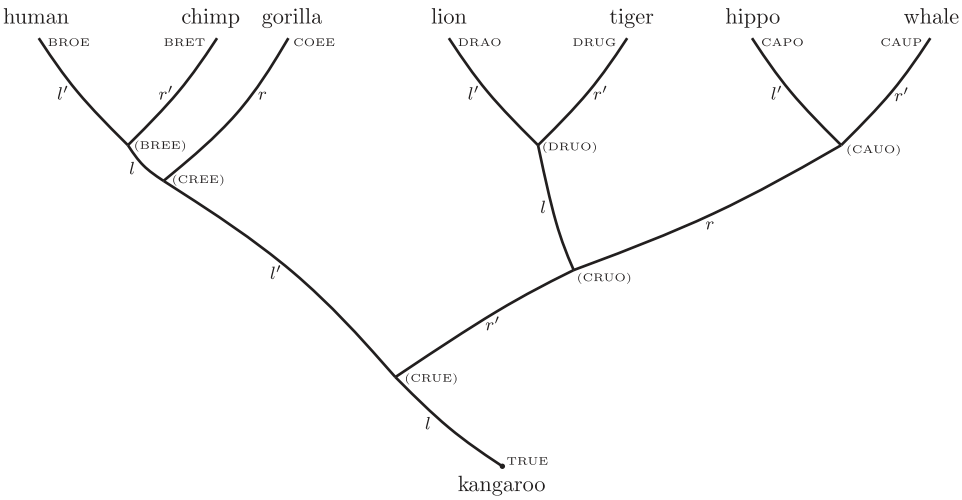

Figure 4. The unique perfect phylogeny T for the four characters described in Table 1. The assignment of ancestral character states is shown in brackets.

A recipe to generate a set S_T is indicated by the letters l, r, l', r' on the edges of the tree. These correspond to alternating left (l, l') and right (r, r') orientations as one moves up the tree under an arbitrary planar embedding. Now, suppose any edge on which l is placed causes a state change for the first character, any edge on which r is placed causes a state change for the second character, and similarly any edge on which l' (respectively, r') is placed causes a state change for the third (respectively, fourth) character. State changes are always to a new state for that character (to ensure the homoplasy-free condition; a state present in one character is free to reappear in a different one). For example, the bottom-most l causes TRUE to change to CRUE. By following this procedure for any binary tree on any number of leaves, it can be shown that S_T satisfies Theorem 1 (for further details, see [**25**]).

When the data are not perfect. The homoplasy-free condition is very strong. A natural relaxation of it, given a character f and a phylogenetic tree T, is to score T by the smallest number of edges of T that need to have differently assigned states at their endpoints in order to extend f to all the vertices of T. This score is called the *parsimony score* of the character on T, denoted, $ps(f, T)$. By the equivalent description of the homoplasy-free condition above, we have $ps(f, T) \geq |f(X)| - 1$, with equality if and only if f is homoplasy-free.

Since there are exponentially many extensions F of f to T, it might be suspected that computing $ps(f, T)$ is hard. However, in 1971, biologist Walter Fitch proposed a

fast algorithm, which was formally verified by mathematician John Hartigan in 1973. This "Fitch–Hartigan algorithm" proceeds via a dynamical programming approach, and it also provides an explicit extension F that minimizes the number of state changes in the tree. For two-state characters, there is also an elegant characterization of the parsimony score using Menger's min-max theorem in graph theory.

Given a sequence of characters on X, a *maximum parsimony tree* for this data is a phylogenetic tree T that minimizes the sum of the parsimony scores of the characters. Finding such a tree can be phrased as a Steiner tree problem in a sequence space, and it turns out to be NP-hard, though branch and bound algorithms exist.

We saw that no sequence of two-state characters shorter than linear in n can give a unique perfect phylogeny. But can we do better if we just want a unique most-parsimonious tree? That is, for each tree $T \in UB(n)$, is there a sequence $\eta(T)$ of two-state characters of length $k = k(n)$ that is sublinear in n and for which T is the unique most-parsimonious tree? A simple counting argument sets an absolute lower bound on k. Let $S(n, k)$ be the set of sequences of two-state characters on $[n]$ of length k. Then k must be at least large enough for the function $T \mapsto \eta(T)$ from $UB(n)$ to $S(n, k)$ to be one-to-one. Since $S(n, k) = 2^{nk}$, this requires that $|UB(n)| \leq 2^{nk}$, which can be rewritten as $k \geq \frac{1}{n} \log_2 |UB(n)|$. If we now invoke (3) and Stirling's approximation for $n!$ to calculate $|UB(n)|$, we see that k must grow at least at the rate $\log(n)$. Remarkably, it was recently shown [**8**] that this primitive logarithmic growth rate can be achieved, and by a function η that can be constructively implemented. Moreover, the parsimony score per character of the resulting sequences $\eta(T)$ on T necessarily tends to infinity as n grows, so this encoding is very far from supporting a perfect phylogeny.

We will return to maximum parsimony in Section 6.

5. METRIC PROPERTIES OF TREES. So far, we have regarded the edges of our trees as being unweighted; however, it is useful—both in biology and in mathematics—to assign weights or lengths to the edges (often called branch lengths in biology). For instance, the length of an edge could correspond to evolutionary time or some measure of the amount of genetic change along that edge. Assigning lengths to edges brings in a further tool to help study and reconstruct trees. It is pivotal to approaches for inferring trees from data that try to estimate an evolutionary distance between pairs of species, as well as for the statistical methods that we will discuss in Section 6.

First, notice that if we have a phylogenetic X-tree T and some function w that assigns strictly positive weights to each edge of the tree, then we can define a metric $d = d_{(T,w)}$ on X by letting $d(x, y)$ be the sum of the weights of the edges on the path in T connecting x and y. When d can be represented by a tree in this way, we say it has a *tree representation* (on T). This leads to two natural questions.

- Does every metric on X have a tree representation?
- Is the choice of T and w in a tree representation unique?

The answers to these questions are no and yes, respectively. Let's consider the first question.

When $|X| = 3$, it is an easy exercise to show that every metric d on X can be represented as a tree metric. But this result is particular to $|X| = 3$ and already runs into problems when $|X| = 4$. It is instructive to see why. Consider the three pairwise sums:

$$d(x, y) + d(w, z), \quad d(x, z) + d(y, w), \quad d(x, w) + d(y, z).$$

If d has a tree representation ($d = d_{(T,w)}$), then two of these pairwise sums must be equal and larger than or equal to the third, regardless of the choice of T. This is illustrated in Figure 5(i). This four-point condition is not usually satisfied by an arbitrary metric d on a set of size four, but when it is, it turns out that d can be represented on a tree. What is much more remarkable is that, for any X, the four point condition holds for all subsets of X of size 4 if *and only if* d has a tree representation. This result, in various forms, dates back to the 1960s.

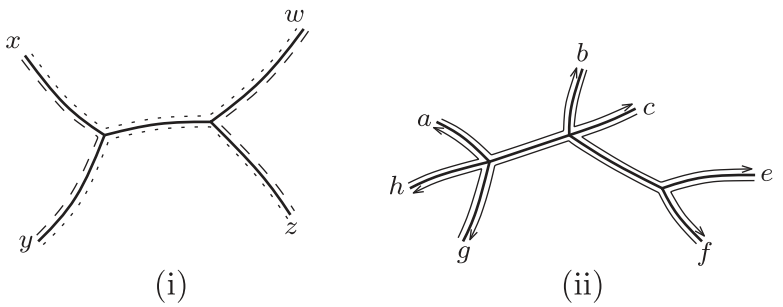

Figure 5. (i): Here $d(x, y) + d(w, z)$ is smaller than $d(x, w) + d(y, z)$ (which, in turn, equals $d(x, z) + d(y, w)$); (ii) a tour of the tree that covers every edge exactly twice.

Consider now the second question: the uniqueness of a tree representation. As before, this question was resolved many decades ago, and uniqueness of both the unrooted tree and the strictly positive edge weights holds; in other words, for trees $T, T' \in U(X)$ and strictly positive edge weightings w and w', we have

$$d_{(T,w)} = d_{(T',w')} \Longrightarrow T = T' \quad \text{and} \quad w = w'. \tag{5}$$

Moreover, to reconstruct a phylogenetic tree with n leaves, we do not usually need all the $\binom{n}{2}$ possible d-values; for a binary tree T, a subset of $2n - 3$ carefully chosen pairs of elements from $[n]$ suffice to uniquely determine both T and w from the value of $d_{(T,w)}$ for those pairs.

A variety of fast (polynomial-time) methods have been devised for building a phylogenetic X-tree from an arbitrary distance function d on X. The most popular, by far, is neighbor joining, and the paper [**40**] that described this heuristic algorithm has now been cited more than 36,000 times. A desirable property of such methods is that when a distance function has a tree representation, then the method will return the underlying tree and edge weights. Moreover, when a distance function δ is sufficiently close to a tree metric $d = d_{(T,w)}$ where T is any binary tree on any number of leaves, many methods also come with a guarantee that they will return T when applied to δ. How close δ needs to be to d depends crucially on $w_{\min}$, the smallest interior edge weight of T; a distance-based tree reconstruction method is said to have *safety radius* r if the method is guaranteed to return the binary tree T when δ differs from $d = d_{(T,w)}$ by less than $r \cdot w_{\min}$ on each pair of leaves. For neighbor-joining, this safety radius is $r = \frac{1}{2}$. It is not hard to show that $\frac{1}{2}$ is the largest possible safety radius for any distance-based tree reconstruction method.

Diversity measures. Given a phylogenetic X-tree T with an edge weighting w, consider $L = \sum_e w(e)$, which is the total sum of the edge weights over the tree. Notice that for the tree in Figure 5(ii) we can write

$$L = \frac{1}{2}[d(a,b) + d(b,c) + d(c,e) + d(e,f) + d(f,g) + d(g,h) + d(h,a)]$$

since the cyclic permutation $(abcefgh)$ traverses the tree in a clockwise order and so covers every edge exactly twice. However, there are other ways to embed the tree in the plane and do this —for example, the cyclic permutation $(aefcbhg)$ also traverses a different planar embedding of this tree in a clockwise order. It is easily shown that for any phylogenetic X-tree, the number N_T of cyclic permutations that traverse the tree in a clockwise order is given by $N_T = \prod_{v \in I(T)} (\deg(v) - 1)!$, where $\deg(v)$ is the degree of vertex v, and $I(T)$ is the set of interior vertices of T [**43**]. For example, for the tree Figure 5(ii)), $N_T = 3! \times 3! \times 2! = 72$. For each cyclic permutation $(x_1, x_2, \ldots, x_n)$, we have $L = \frac{1}{2}[d(x_1, x_2) + d(x_2, x_3) + \cdots + d(x_{n-1}, x_n) + d(x_n, x_1)]$. If we average these expressions for L over all the N_T cyclic permutations that traverse T in clockwise order, then it is clear that we can write

$$L = \sum_{\{x,y\}} \lambda_T(x,y) d(x,y) \tag{6}$$

for some non-negative coefficients $\lambda_T(x,y)$. These coefficients can be easily described in terms of the number and degrees of the vertices in T on the path between x and y. If T is a binary tree, then $\lambda_T(x,y) = (\frac{1}{2})^{|I(x,y)|}$, where $I(x,y)$ is the set of interior vertices in the path between x and y in T. For instance, in the case of the quartet tree in Figure 5(i), this gives

$$L = \frac{1}{2}d(x,y) + \frac{1}{2}d(w,z) + \frac{1}{4}(d(x,w) + d(x,z) + d(y,w) + d(y,z)).$$

More generally, for any phylogenetic tree T, it can be shown that

$$\lambda_T(x,y) = \prod_{v \in I(x,y)} (\deg(v) - 1)^{-1}.$$

The identity (6) suggests a new way to build phylogenetic trees from distances, which is called Balanced Minimum Evolution (BME) [**36**]. Given an arbitrary distance function (not necessarily a tree metric) δ on X, this method scores each phylogenetic X-tree T by the value $L_\delta(T) = \sum_{\{x,y\}} \lambda_T(x,y)\delta(x,y)$ and searches for a tree T that has the smallest $L_\delta(T)$ score. If δ has a tree representation on some tree T, then this tree has the smallest L_δ score; moreover, like neighbor-joining, BME has the largest possible safety radius of $\frac{1}{2}$. BME can be viewed algebraically as a type of "weighted least squares" method [**14**]. BME is also closely connected with the neighbor-joining method, both algorithmically [**21**], and via polyhedra geometry [**23**].

As well as considering the total diversity of the tree L, we can also consider how much diversity is spanned by different subsets of leaves. This measure is called *phylogenetic diversity* (PD) and is relevant to biodiversity conservation [**37**]. Formally, given a phylogenetic X-tree T and a positive edge weighting w, we can associate to each subset Y of X a non-negative value, denoted $PD(Y)$, equal to the sum of the weights of the edges of the minimal subtree of T that connect the leaves in Y. For example, $L = PD(X)$, and $d_{(T,w)}(x,y) = PD(\{x,y\})$. Just as the PD scores of subsets of size $k = 2$ (i.e., distances) can be used to reconstruct a tree so can the PD scores of subsets of size k for any k up to (but not exceeding) $\lceil n/2 \rceil$ [**35**].

The function PD is clearly monotone—the PD of a set is always greater than the PD of any strict subset; moreover, PD enjoys a strong exchange property: For any

subset Y_1 of X of size at least two and any subset Y_2 of X that is larger in size than Y_1, there always exists an element $y \in Y_2 - Y_1$ for which

$$PD(Y_1 \cup \{y\}) + PD(Y_2 - \{y\}) \geq PD(Y_1) + PD(Y_2).$$

This property formally justifies a simple and fast strategy for finding a subset Y of X of any given size k having maximal PD for sets of that size. The strategy is simply the greedy one: First select two leaves x, y that are furthest apart in the tree (i.e., maximize $d_{(T,w)}(x, y)$) and then sequentially add a leaf that increases the PD score by the maximum amount to the tree so-far constructed until k leaves are present. The collection of subsets of X that have maximal PD score for their cardinality form what is known in combinatorics as a greedoid.

A more sophisticated mathematical approach to the study of distances is T-theory (tight-span), pioneered by Andreas Dress and colleagues [**12**], and extended recently to the diversity setting [**7**].

Distances and diversities can also be generalised to allow the edge weights to take nonzero values in an arbitrary Abelian group $\mathcal{G}$ (e.g., the special case $\mathcal{G} = (\mathbb{R}, +)$ allows negative edge weights). Several of the main results above extend with minor modification. There is one fly in the ointment, however—for distances, problems arise if $\mathcal{G}$ has elements of order 2 (for instance, uniqueness of the tree representation fails; this is apparent from the 15 phylogenetic trees having the shape shown in Figure 2(e) with edges assigned the element 1 of $\mathcal{G} = (\{0, 1\}, +)$ that induce exactly the same distance function). But uniqueness can be restored by moving from distances to diversities, where not just pairs but also triples of leaves are considered [**13**].

6. MARKOV MODELS AND THE FELSENSTEIN ZONE. A major advance in phylogenetics has been the development of stochastic models to describe the evolution of genetic sequences and genomes on a tree. For genetic sequences, these models typically describe point substitutions that occur at sites in the DNA sequence that codes for some particular gene. Such models allow biologists to convert the sequences we observe today at the leaves of the tree into an estimate of the tree itself (and perhaps its branch lengths, or ancestral states within the tree). By combining these gene trees one can in turn estimate the species tree.

The rise of statistical phylogenetics was pioneered in the 1960s and 1970s by Anthony Edwards, Joseph Felsenstein, and others (including David Sankoff, with a visionary paper in this journal [**41**]). Today's methods of choice are based on maximum likelihood and Bayesian approaches. Stochastic models assume that characters evolve independently on a tree, and the evolution of each character is described by some Markovian process; this may be the same across the characters or vary (for instance, some characters may evolve more rapidly than others).

One of the catalysts that ushered in this stochastic approach was a landmark 1978 paper by Joseph Felsenstein [**20**]. He showed that if characters evolve independently under a simple stochastic process, then existing methods like maximum parsimony (discussed above) can be seriously misled. So, as the number of characters increases, it would be increasingly certain that the maximum parsimony tree will be a different tree from the true tree (i.e., the one on which the characters evolved). By contrast, other methods (like maximum likelihood) are, under certain conditions, provably statistically consistent and so converge on the true tree as the number of characters grows.

Felsenstein considered a simple process involving just two states—let's call them α and β—which can flip between states with equal probability. This process is familiar

in coding theory as the binary symmetric channel. In phylogenetics, we apply this process to the edges of a tree—each edge e of the tree has a certain probability p_e of a change of state between its endpoints, and, as in coding theory, it is assumed that p_e lies strictly between 0 and 0.5. The model also assumes that the (marginal) state at any given leaf is uniform (i.e., no state is preferred) and that changes of states on different edges are independent events.

Felsenstein's tree is shown in Figure 6(b)—we can imagine it as a tree in which there has been an accelerated rate of evolution (resulting in higher probabilities of change) in two nonadjacent lineages. It can also be realized on a rooted tree as in Figure 6(a), with a single rate increase in one short branch (the branch leading to 1) and a distant out-group species (4). Denote the probabilities of change on the edges of the tree in Figure 6(b) by the values $p_1, \ldots, p_5$, as shown.

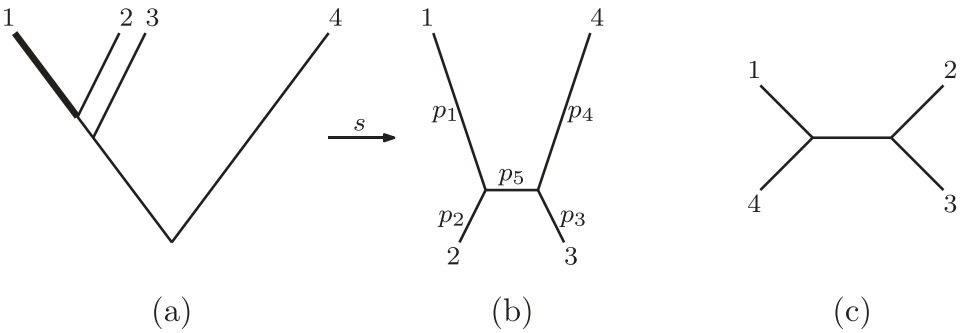

Figure 6. (a) A high rate of evolution on the lineage leading to species 1 and a distant out-group species (4) can be modelled by a Markov process on the associated unrooted tree (obtained by suppressing the root) in (b); for this tree T, if p_1 and p_4 are large enough relative to the other p_i values, the maximum parsimony tree for a large number of characters generated on T is likely to be the tree T' shown in (c).

Now, there are 2^n different ways to assign the two states to a set X of size n, but if we identify complementary assignments, obtained by interchanging α and β (these two assignments have equal probability under the model) we get just 2^{n-1} distinct *patterns*. For a subset A of $[n-1]$, let p_A be the probability of generating a pattern at the leaves of the tree in which A is precisely the leaves that are in different state to leaf n. For example, $p_\emptyset$ is the probability that all leaves are in the same state (i.e., all α or all β). For the tree in Figure 6(b), a check of the four possible pairs of states at the two interior vertices gives

$$p_\emptyset = (1-p_1)(1-p_2)(1-p_3)(1-p_4)(1-p_5) + p_1p_2p_3p_4(1-p_5) + p_1p_2p_5(1-p_3)(1-p_4) + p_3p_4p_5(1-p_1)(1-p_2).$$

There are various ways to compute the p_A values, but one particularly elegant way that holds for any phylogenetic tree with n leaves is by the following identity:

$$p_A = \frac{1}{2^{n-1}} \sum_{\substack{B \subseteq [n] \\ |B| \equiv 0 \bmod 2}} (-1)^{|A \cap B|} \prod_{e \in P(T,B)} (1-2p_e), \tag{7}$$

where $P(T, B)$ is the unique set of edges of T that lie on any set of edge-disjoint paths in T that connect pairs of leaves in the even cardinality set B. For the tree in Figure 6(b), if we let $x_i = (1 - 2p_i)$ and take $A = \emptyset$ (so that $(-1)^{|A \cap B|} = 1$ for all B in Eqn. (7)), then we obtain

$$p_\emptyset = \frac{1}{8}(1 + x_1x_2 + x_3x_4 + x_1x_3x_5 + x_2x_3x_5 + x_1x_4x_5 + x_2x_4x_5 + x_1x_2x_3x_4). \quad (8)$$

All other p_A values are obtained from the right-hand side of (8) by replacing $+$ by $-$ for exactly half the terms. The somewhat mysterious representation in (7) follows from a combinatorial study of this model (in which $[(-1)^{|A\cap B|}]$ turns out to be a Hadamard matrix) due to Mike Hendy [**24**] and generalized to other models using discrete Fourier analysis by Evans and Speed [**18**], and Székely *et al.* [**48**].

With this in hand, we can now establish the main ingredient in Felsenstein's classic result for maximum parsimony.

Theorem 2. *For a character generated on tree T in Figure 6(b) under the two-state symmetric model with $p_1 = p_4 = P$ and $p_2 = p_3 = p_5 = Q$, the expected parsimony score of T is larger than for the tree T' in Figure 6(c) precisely when $P^2 > Q(1 - Q)$.*

Proof. The only two-state characters that have different parsimony scores on T and T' are those that correspond to patterns that we will denote by f_{12} and f_{23}, where $f_{12}(1) = f_{12}(2) \neq f_{12}(3) = f_{12}(4)$ and $f_{23}(2) = f_{23}(3) \neq f_{23}(1) = f_{23}(4)$. Notice that a character of type f_{12} has a parsimony score of 1 on T and 2 on T', while a character of type f_{23} has a parsimony score of 1 on T' and 2 on T. Moreover, under the two-state symmetric model on T, the probabilities of generating the patterns f_{12} and f_{23} are p_{12} and p_{23}, respectively. Now, for a character generated by this model on T, consider the random variable Δ that is the parsimony score of that character on T minus the parsimony score of that character on T'. Then the expected value of Δ, denoted $\mathbb{E}[\Delta]$, satisfies

$$\mathbb{E}[\Delta] = p_{23} - p_{12}. \quad (9)$$

Applying (7) for $n = 4$, letting $x_i = (1 - 2p_i)$, and with $A = \{1, 2\}$ and $\{2, 3\}$:

$$p_{12} = \frac{1}{8}(1 + x_1x_2 + x_3x_4 - x_1x_3x_5 - x_2x_3x_5 - x_1x_4x_5 - x_2x_4x_5 + x_1x_2x_3x_4),$$

and

$$p_{23} = \frac{1}{8}(1 - x_1x_2 - x_3x_4 - x_1x_3x_5 + x_2x_3x_5 + x_1x_4x_5 - x_2x_4x_5 + x_1x_2x_3x_4).$$

Substituting these identities into (9) gives $\mathbb{E}[\Delta] = \frac{1}{4}(-x_1x_2 - x_3x_4 + x_2x_3x_5 + x_1x_4x_5)$. Now, setting $x_1 = x_4 = u = (1 - 2P)$ and $x_2 = x_3 = x_5 = v = (1 - 2Q)$ we obtain

$$\mathbb{E}[\Delta] = \frac{v}{4}[u^2 + v^2 - 2u] = v(P^2 - Q(1 - Q)),$$

and so $\mathbb{E}[\Delta] > 0$ precisely if $P^2 > Q(1 - Q)$. This completes the proof. ∎

Theorem 2, together with the law of large numbers (or the central limit theorem), ensures that for k characters generated by the T (with these p_i values), a different tree, namely T', will have a lower parsimony score than T, with probability converging to 1 as k grows. Intuitively, parallel changes on the two long branches of T become more probable than a single change on the short edges. So, through the eyes of parsimony, it is more optimal to join these two edges together in the reconstruction. This phenomenon of "long branch attraction" has been observed in biological data [**26**].

While parsimony can fail to recover the true tree, there are statistically consistent methods for inferring it. A particularly simple one for the two-state symmetric model relies on the following distance function on X. For $x, y \in X$, let

$$\hat{\mu}(x, y) = -\frac{1}{2}\log(1 - 2\hat{p}(x, y)),$$

where $\hat{p}(x, y)$ is the proportion of characters that assign different states to x and y. Then provided we apply a distance-based tree reconstruction method with a positive safety radius (*c.f.* Section 5)—we are guaranteed to recover the underlying (unrooted) tree from k independently evolved characters, as k grows. The reason is that, as $k \to \infty$, the law of large numbers ensures that $\hat{p}(x, y)$ will converge to the probability $p(x, y)$ that leaves x and y are in different states, and so $\hat{\mu}(x, y)$ converges to $\mu(x, y) = -\frac{1}{2}\log(1 - 2p(x, y))$. It is then an easy exercise to show that μ has a tree representation on the true tree T with the edge weighting $w(e) = -\frac{1}{2}\log(1 - 2p_e)$. That is, $\mu = d_{(T,w)}$. The uniqueness result (i.e., the implication in (5)) then ensures the reconstruction of both the unrooted tree and the edge weights from μ (and thereby $\hat{\mu}$ for k sufficiently large).

Biologists deal with much more complex models of character evolution than the two-state symmetric model, often on 4, 20, or 64 states (corresponding to DNA, amino acid, and codon sequences, respectively). For a general Markov model involving any state space, there are ways to construct a metric that has a tree representation on T, based on the logarithm of the determinant of the matrix of the joint probabilities of states for each pair of species. In this way, the tree is identifiable from the probability distribution of characters. This identifiability result is enough to ensure that methods like maximum likelihood are statistically consistent. However, for mixtures of such processes, the identifiability of the tree can easily be lost (mixtures of Markov processes are generally no longer Markovian). This can be important for biologists—if there are too many parameters to estimate from the data, then one may lose the ability to infer the one(s) we are interested in (such as the tree). A striking example of this loss was provided for the two-state symmetric model [**31**]: If 50% of DNA sites evolve on a four-species tree T with one carefully chosen set of branch lengths and 50% evolve on the same tree under a different chosen collection of branch lengths, then the expected proportion of site patterns is exactly identical to that in which all sites evolve on a different tree with appropriately chosen edge lengths. More recent results appear in [**4, 39**].

To obtain a deeper understanding of Markov processes on trees, techniques from commutative algebra and Lie algebra theory have proved invaluable [**2, 46, 47**]. In particular, these techniques can be applied to determine the extent to which trees and other parameters of the model can be reconstructed from data (the identifiability issue mentioned above) [**3**], a topic that is part of a broader emerging area called algebraic statistics [**15**]. The combinatorial topology and geometry of two different notions of tree space are also of interest [**5, 33**], as is the question of how much data we need to reconstruct a tree accurately [**9**].

7. CURRENT CHALLENGES. We have provided a brief overview of some of the central ideas in phylogenetics, but much has been omitted and the reader interested in this area may wish to consult [**19, 44**] for further details.

Two areas that are currently very active and where mathematical and computational approaches play a key role include the following.

- Using probability theory and combinatorics to study how the genealogy of each gene (the gene tree) for a set of species relates to the species' phylogenetic tree (the species tree). Biologists typically now have very large numbers (thousands) of gene trees by which to compare species, but these trees can differ from the species tree by a process called "incomplete lineage sorting." By considering how genes trace back in time and coalesce, it is possible to explain gene tree discordance and predict species trees from these conflicting gene trees (see, e.g., [**3, 11, 28, 32**]).
- Extending phylogenetic tree theory to phylogenetic networks, which are graphs that either display uncertainty in the data as to the likely species tree (implicit networks) or which provide an explicit representation of evolution where there has been reticulation (such as the formation of hybrid species (see, for example, [**27**]). The patchy distribution of genes across taxa and lateral gene transfer also lead to further combinatorial and computational challenges [**38**].

Finally, we have seen how any phylogenetic X-tree can be encoded by its associated set of splits and also by the leaf-to-leaf distances the tree induces under an edge weighting. However, there is a third encoding, obtained by considering the quartet trees that are induced by the tree on subsets of X of size four. This association has led to some of the deepest results in phylogenetics (see, e.g., [**22**]) and the exploration of the links between these three equivalent ways of encoding phylogenetic trees forms the basis of the emerging area of phylogenetic combinatorics (see [**12**] for more details).

ACKNOWLEDGMENTS. Funding for this work was made possible by the NZ Marsden Fund and the Allan Wilson Centre. I thank Simone Linz, Elliott Sober, Amelia Taylor, and three anonymous reviewers for several helpful comments.

REFERENCES

1. D. J. Aldous, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, *Stat. Sci.* **16** (2001) 23–34.
2. E. S. Allman, J. A. Rhodes, Phylogenetic ideals and varieties for the general Markov model, *Adv. Appl. Math.* **40** (2008) 127–148, http://dx.doi.org/10.1016/j.aam.2006.10.002.
3. E. S. Allman, J. H. Degnan, J. A. Rhodes, Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent, *J. Math. Biol.* **62** (2011) 833–862.
4. E. S. Allman, J. A. Rhodes, S. Sullivant, When do phylogenetic mixture models mimic other phylogenetic models?, *Syst. Biol.* **61** (2012) 1049–1059, http://dx.doi.org/10.1093/sysbio/sys064.
5. L. J. Billera, S. P. Holmes, K. Vogtmann, Geometry of the space of phylogenetic trees, *Adv. Appl. Math.* **27** (2001) 733–767, http://dx.doi.org/10.1006/aama.2001.0759.
6. P. Bonizzoni, A. P. Carrieri, G. Della Vedova, R. Dondi, T. M. Przytycka, When and how the perfect phylogeny model explains evolution, in *Discrete and Topological Models in Molecular Biology*. Edited by N. Jonoska, M. Saito. *Natural Computing Series*, Springer-Verlag Berlin, Heidelberg, 2014, 67–83.
7. D. Bryant, P. F. Tupper, Hyperconvexity and tight span theory for diversities, *Adv. Appl. Math.* **231** (2012) 3172–3198.
8. J. Chai, E. A. Housworth, On the number of binary characters needed to recover a phylogeny using maximum parsimony, *Bull. Math. Biol.* **73** (2011) 1398–1411, http://dx.doi.org/10.1007/s11538-010-9579-3.
9. C. Daskalakis, E. Mossel, S. Roch, Evolutionary trees and the Ising model on the Bethe lattice: A proof of Steel's conjecture, *Probab. Theor. Relat. Fields* **149** (2011) 149–189.
10. W. H. E. Day, F. R. McMorris, *Axiomatic Consensus Theory in Group Choice and Biomathematics.* SIAM Frontiers in Applied Mathematics, SIAM, Philadelphia PA, 2003.

11. J. H. Degnan, N. A. Rosenberg, T. Stadler, The probability distribution of ranked gene trees on a species tree, *Math. Biosci.* **235** (2012) 45–55, http://dx.doi.org/10.1016/j.mbs.2011.10.006.
12. A. Dress, K. T. Huber, J. Koolen, V. Moulton, A. Spillner, *Basic Phylogenetic Combinatorics.* Cambridge Univ. Press, New York, 2012, http://dx.doi.org/10.1017/cbo9781139019767.
13. A. Dress, M. Steel, Phylogenetic diversity over an Abelian group, *Ann. Combin.* **11** (2007) 143–160.
14. R. Desper, O. Gascuel, Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting, *Mol. Biol. Evol.* **21** (2004) 587–598.
15. M. Drton, B. Sturmfels, S. Sullivant, *Lectures on Algebraic Statistics.* Birkhäuser, Berlin, 2009.
16. J. Edmonds, R. Giles, A min-max relation for submodular functions on graphs, Studies in Integer Programming, Proceedings of Workshop on Programming, Bonn, 1975, *Ann. Discrete Math.* **1** (1977) 185–204.
17. P. L. Erdös, L. A. Székeley, Applications of antilexicographic order. I. An enumerative theory of trees, *Adv. Appl. Math.* **10** (1989) 488–496.
18. S. N. Evans, T. P. Speed, Invariants of some probability models used in phylogenetic inference, *Ann. Stat.* **21** (1993) 355–377.
19. J. Felsenstein, *Inferring Phylogenies.* Sinauer Associates, Sunderland, MA, 2004.
20. J. Felsenstein, Cases in which parsimony or compatibility methods will be positively misleading, *Syst. Zool.* **27** (1978) 401–410.
21. O. Gascuel, M. Steel, Neighbor-joining revealed, *Mol. Biol. Evol.* **23** (2006) 1997–2000.
22. S. Grünewald, Slim sets of binary trees, *J. Comb. Theory A* **119** (2012) 323–330, http://dx.doi.org/10.1016/j.jcta.2011.09.007.
23. D. C. Haws, T. L. Hodge, R. Yoshida, Optimality of the neighbor joining algorithm and faces of the balanced minimum evolution polytope, *Bull. Math. Biol.* **73** (2011) 2627–2648.
24. M. D. Hendy, The relationship between simple evolutionary tree models and observable sequence data, *Syst. Biol.* **38** (1989) 310–321.
25. K. Huber, V. Moulton, M. Steel, Four characters suffice to convexly define a phylogenetic tree, *SIAM J. Discrete Math.* **18** (2005) 835–843.
26. J. Huelsenbeck, Is the Felsenstein zone a fly trap? *Syst. Biol.* **46** (1997) 69–74.
27. D. H. Huson, R. Rupp, C. Scornavacca, *Phylogenetic Networks: Concepts, Algorithms and Applications.* Cambridge Univ. Press, Cambridge, UK, 2011.
28. L. L. Knowles, L. S. Kubatko, *Estimating Species Trees: Practical and Theoretical Aspects.* Wiley-Blackwell, Hoboken, NJ, 2010.
29. A. Lambert, T. Stadler, Birth-death models and coalescent point processes: The shape and probability of reconstructed phylogenies, *Theor. Popul. Biol.* **90** (2013) 113–128, http://dx.doi.org/10.1016/j.tpb.2013.10.002.
30. C. Linnaeus, *Systema Naturae.* First edition. 1735.
31. F. A. Matsen, M. Steel, Phylogenetic mixtures on a single tree can mimic a tree of another topology, *Syst. Biol.* **56** (2007) 767–775.
32. E. Mossel, S. Roch, Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci, *IEEE/ACM Trans. Comput. Biol. Bioinf.* **7** (2010) 166–171, http://dx.doi.org/10.1109/tcbb.2008.66.
33. V. Moulton, M. Steel, Peeling phylogenetic "oranges," *Adv. Appl. Math.* **33** (2004) 710–727.
34. M. Nikaido, A. P. Rooney, N. Okada, Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: Hippopotamuses are the closest extant relatives of whales, *Proc. Natl. Acad. Sci. USA* **96** (1999) 10261–10266.
35. L. Pachter, D. Speyer, Reconstructing trees from subtree weights, *Appl. Math. Lett.* **17** (2004) 615–621.
36. Y. Pauplin, Direct calculation of a tree length using a distance matrix, *J. Mol. Evol.* **51** (2000) 41–47.
37. A. Purvis, P.-M. Agapow, J. L. Gittleman, G. M. Mace, Nonrandom extinction and the loss of evolutionary history, *Science* **288** (2000) 328–330.
38. S. Roch, S. Snir, Recovering the tree-like trend of evolution despite extensive lateral genetic transfer: A probabilistic analysis, *J. Comput. Biol.* **20** (2013) 93–112, http://dx.doi.org/10.1089/cmb.2012.0234.
39. J. A. Rhodes, S. Sullivant, Identifiability of large phylogenetic mixture models, *Bull. Math. Biol.* **74** (2012) 212–231, http://dx.doi.org/10.1007/s11538-011-9672-2.
40. N. Saitou, M. Nei, The neighbor-joining method: A new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.* **4** (1987) 406–425.
41. D. Sankoff, Reconstructing the history and geography of an evolutionary tree, *Amer. Math. Monthly* **79** (1972) 596–603.
42. E. Schröder, Vier combinatorische probleme, *Z. Angew. Math. Phys.* **15** (1870) 361–376.

43. C. Semple, M. Steel, Cyclic permutations and evolutionary trees, *Adv. Appl. Math.* **32** (2004) 669–680.
44. C. Semple, M. Steel, *Phylogenetics.* Oxford Univ. Press, 2003.
45. B. Shutters, S. Vakati, D. Fernández-Baca, Incompatible quartets, triplets, and characters, *Algorithms Mol. Biol.* **8** (2013) 11, http://dx.doi.org/10.1186/1748-7188-8-11.
46. J. G. Sumner, B. R. Holland, P. D. Jarvis, The algebra of the general Markov model on phylogenetic trees and networks, *Bull. Math. Biol.* **74** (2012) 858–880, http://dx.doi.org/10.1007/s11538-011-9691-z.
47. J. G. Sumner, J. Fernández-Sánchez, P. D. Jarvis, Lie Markov models, *J. Theor. Biol.* **298** (2012) 16–31, http://dx.doi.org/10.1016/j.jtbi.2011.12.017.
48. L. A. Székely, M. A. Steel, P. L. Erdös, Fourier calculus on evolutionary trees, *Adv. Appl. Math.* **14** (1993) 200–216.

MIKE STEEL is director of the Biomathematics Research Centre at University of Canterbury, New Zealand, where he teaches mathematics and statistics. He is a fellow of the Royal Society of New Zealand and deputy director of the Allan Wilson Centre. When not doing mathematics, he likes to get out for long runs in the mountains.
School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
mike.steel@canterbury.ac.nz