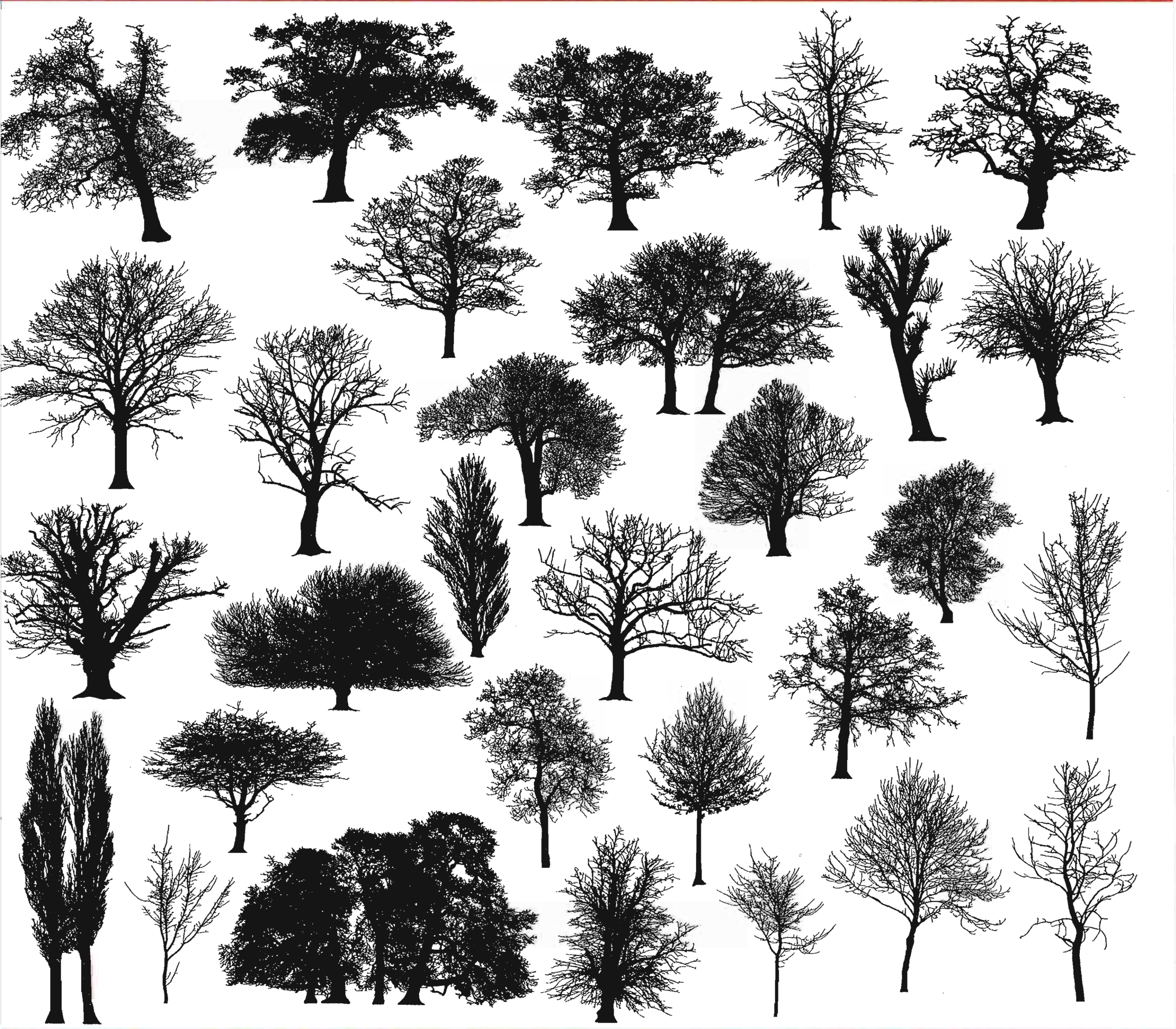
FEBRUARY 2009

math HORIZONS

PUBLISHED BY THE MATHEMATICAL ASSOCIATION OF AMERICA



Searching for the Tree of Life...

"Searching for the 'true tree' for twenty species is like trying to find a needle in very large haystack. Yet biologists are now routinely attempting to build trees with hundreds and sometimes thousands of species."

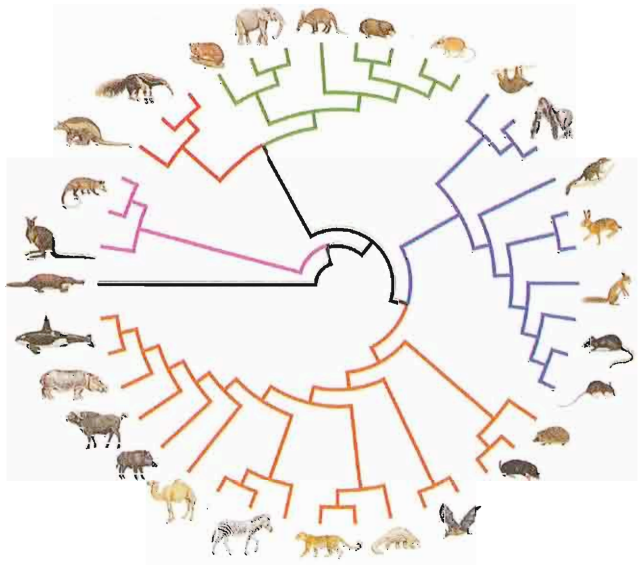
Mathematical Aspects of the 'Tree of Life'

Charles Semple and Mike Steel
University of Canterbury, New Zealand

Trees have long been used for illustrating certain phenomena in nature. They describe the structure of braided rivers, classify acyclic hydrocarbons in chemistry, and model the growth of cell division in physiology. In evolutionary biology, trees describe how species evolved from a common ancestor. Evolutionary trees date back (at least) to Charles Darwin, who made an early sketch of one in a notebook from 1837, more than 20 years before his *Origin of Species*. In Darwin's day, the main evidence available for reconstructing such trees, apart from some fossils, was morphology and physiology—the physical details of different species (does an animal have wings or not, how many petals does a plant have, and so forth). As Darwin wrote in 1872:

> *We possess no pedigrees or amorial bearings; and we have to discover and trace the many diverging lines of descent in our natural genealogies, by characters of any kind which have long been inherited.*

The problem with morphological and fossil data is that they are patchy, and can be misleading. Also the evolution of morphological characters is often complex and difficult to model. Today, biologists prefer genetic data to reconstruct and study evolutionary trees, using methods that increasingly are based on mathematical ideas. This field is known as *phylogenetics*—*phylo* means race or tribe, and thus phylogenetics is the study of grouping together species with similar genetic characteristics. Since it attempts to reconstruct the distant past, it is usually impossible to check directly whether an inferred tree is correct or not. So phylogenetics is somewhat akin to a field like forensic science that aims to reconstruct a crime scene for which no witness was present. Phylogenetic methods are used to study other problems in biology—such as how different strains of a virus (for example, of HIV or influenza) are related, how much biodiversity is threatened by extinction, and how populations, such as modern humans, came to be distributed across the planet. Phylogenetic techniques have even been used to study the evolution of languages and the copying history of old manuscripts.



A phylogenetic (evolutionary) tree of 42 representative mammals reconstructed using a Bayesian approach by concatenating 12 protein coding mitochondrial genes. The main mammalian groups are colour-coded (Frédéric Delsuc and Nicolas Lartillot, unpublished).

In this article, we will describe some of the mathematical aspects of phylogenetics, starting from the elementary principles, but also describing one or two deeper and more recent results. Definitions of some basic combinatorial terms used in this article are provided in a box on page 8.

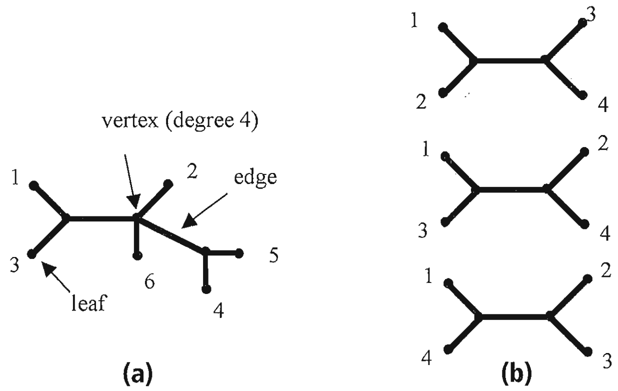
Counting Trees

One of the most famous enumeration formulae in combinatorics is Arthur Cayley's formula from 1889, for the number of trees with n labeled vertices. This number is n^{n-2}, a simple formula that has many known proofs, most of which are surprisingly complicated.

In phylogenetics, we are interested in counting a different class of trees, called phylogenetic (evolutionary) trees. A *phylogenetic tree* is a tree whose leaves are labeled, but the remaining vertices are unlabeled and of degree at least three.

The leaves correspond to the species we see today, while the unlabeled vertices correspond to unknown ancestral species. The most 'informative' type of phylogenetic trees are those that have the largest number of edges possible for the number of leaves—these are sometimes called *fully resolved* or *binary* phylogenetic trees—and this is the set of trees we will count. Every vertex in a binary philogenetic tree that is not a leaf must have degree 3, for otherwise two edges incident with such a vertex v could be plucked from v and joined together at a new vertex, and a single new edge could be placed with one end at the new vertex and the other end at v, resulting in a phylogenetic tree with the same leaves as the original, and one more edge.

Even though the unlabeled vertices of a binary phylogenetic tree all have degree 3, the reason they are called 'binary' is because if we direct all the edges away from some leaf (which might represent some 'outgroup' species that is distantly related to the other species), then each non-leaf vertex encountered as one traverses the tree can be viewed as a speciation event—where each species splits into two daughter species. These concepts are illustrated below.



(a) A phylogenetic tree on six leaves. **(b)** The three possible binary phylogenetic trees on four leaves.

The observant reader may have noticed that in the above definition, a phylogenetic tree has no root (representing a common ancestor). In practice, most reconstruction methods build an unrooted phylogenetic tree. The often controversial decision of where to place the root is done later.

To count the number of binary phylogenetic trees, we may assume that the leaves are labeled $1, 2, \ldots, n$. There are three unlabeled binary trees for $n = 4$, shown above. How many binary phylogenetic trees are possible on n leaves? This problem was solved long ago—in fact, 21 years after Darwin published his *Origin of Species*, the mathematician Ernst Schröder was counting classes of trees for quite a different purpose, and these trees included what we now call binary phylogenetic trees.

Let $N(n)$ be the number of binary phylogenetic trees on the leaf set $\{1, 2, \ldots, n\}$. To determine $N(n)$, we can use a well-known and elementary result from graph theory called the *Handshaking Lemma*: the sum of the degrees of the vertices of any graph is equal to twice the number of edges. For a binary phylogenetic tree, if we denote the number of unlabeled (non-leaf) vertices by I and the number of edges by E, then the Handshaking Lemma gives:

$$3I + n = 2E. \qquad (1)$$

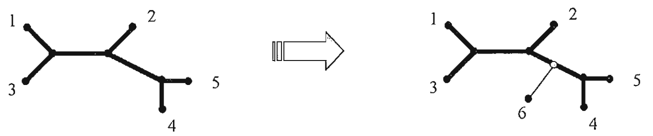
But a tree has one more vertex than its number of edges (see box), so we also have:

$$E = n + I - 1. \qquad (2)$$

Combining Equations (1) and (2) gives $3I + n = 2(n + I - 1)$, and so $I = n - 2$. Since, by (2), $E = n + I - 1$, the number of edges in a binary phylogenetic tree is

$$E = 2n - 3, \qquad (3)$$

an equation that one could also prove by induction on n. Curiously, Equation (3) does not depend on the shape of the tree, just the number of leaves. To determine $N(n)$, we need another insight—each binary phylogenetic tree T on leaf set $\{1, 2, \ldots, n\}$ is obtained from a binary phylogenetic tree T' on leaf set $\{1, 2, \ldots, n-1\}$ by subdividing an edge and attaching leaf n by a new edge to the vertex resulting from this subdivision as shown.



Subdividing an edge to create a phylogenetic tree with 6 leaves.

Different choices of T' or edges to subdivide produce different choices for T and, as each T' has $(2n - 5)$ edges (by (3)), we have:

$$N(n) = (2n - 5)\, N(n - 1). \qquad (4)$$

This recurrence, together with the boundary condition $N(3) = 1$, implies that

$$N(n) = (2n-5)\times(2n-7)\times\cdots\times 3\times 1. \qquad (5)$$

That is, $N(n)$ is the product of the first $n-2$ odd numbers. Written $(2n-5)!!$, this number is sometimes called a 'semi-factorial'. It can also be written as a ratio of factorials and powers of 2 as follows:

$$\frac{(2n-4)!}{(n-2)!2^{n-2}} \qquad (6)$$

This number shows up in other quite different contexts—for example, it is the number of ways that $2n - 4$ people can be paired up into $n - 2$ teams of two. Finding a way to match a binary phylogenetic tree to each such pairing is not easy, but possible.

How large is $N(n)$? Well, for 10 species we have $N(10) = 2{,}027{,}025$; for $n = 20$, we have $N(20) = 2\times10^{20}$, prompting the biologist Walter Fitch to remark some years ago that:

> *We have, for twenty species, more than a gram molecular weight of evolutionary tree!*

Searching for the 'true tree' for twenty species is like trying to find a needle in very large haystack. Yet biologists are now routinely attempting to build trees with hundreds and sometimes thousands of species. To realize the smallness of the needle, consider how many millenia it would take a computer that can check a million trees per second to search through $N(20)$ trees.

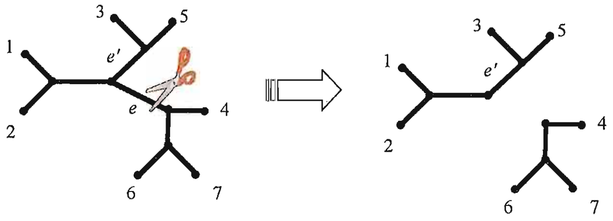
This leads to a very natural question—is there enough DNA sequence data available to reconstruct accurately the one-in-a-zillion tree that describes the evolution of the species you are interested in? Or are the trees being reconstructed almost certain to be false in some details? We can give a crude counting argument to provide a lower bound on the genome length k required to accurately reconstruct binary phylogenetic trees on n species. The number of genomes of length k is 4^k (since a genome is an ordered sequence whose alphabet consists of the four bases of DNA – A, C, G and T). Representing each species by its genome, the number of data sets consisting of n genomes of length k is therefore $(4^k)^n = 4^{nk}$. Now, suppose we have a method that can reconstruct each binary phylogenetic tree from some possible data set. Think of this method as a function from the set A of data sets, each consisting of n genomes (one for each species) of length k, to the set B of binary phylogenetic trees on n species. Then the ability to reconstruct each possible tree means that we want this function to be onto, and so $|B| \leq |A|$. In other words:

$$N(n) \leq 4^{nk}. \qquad (7)$$

Some straightforward analysis (applying Stirling's approximation for $n!$ to (6) in (7)) shows that k must grow at least at the rate $\log(n)$. Of course, this bound is very crude, and it would be useful to derive an upper bound on k under some model of genome evolution, and see how different to $\log(n)$ it is. We will describe the surprising answer to this problem at the end of this article, but first we need to say more about mathematical aspects of phylogenetic trees and how they can describe data.

Properties of Phylogenetic Trees and Data

Phylogenetic trees can be described in many ways—one of the simplest and most versatile is via its splits. Imagine taking a pair of scissors and snipping an edge of a phylogenetic tree. This disconnects the tree into two components, and the leaves in each component form a bipartition, or *split* of the leaf set $\{1, 2, \ldots, n\}$. For example, cutting the edge e in the tree shown below results in the two sets $A = \{1, 2, 3, 5\}$ and $B = \{4, 6, 7\}$. We denote such a split by writing $A|B$ (or, equivalently, $B|A$). The set of splits that are generated from a tree in this way carry enough information that one can always uniquely recover the tree just from its splits. Indeed, there is an easy and fast algorithm for doing this called *tree-popping*, developed in 1983 by a biologist Christopher Meacham.



Cutting an edge e produces the split $\{1,2,3,5\}|\{4,6,7\}$. Cutting both e and e' produces the partition $\{\{1,2\}, \{3,5\}, \{4,6,7\}\}$.

However, not every set of splits of $\{1, 2, \ldots, n\}$ corresponds to the splits of a phylogenetic tree with leaf set $\{1, 2, \ldots, n\}$—the precise conditions for this were found by Peter Buneman in 1971. Buneman was studying trees for a quite different purpose, namely reconstructing the copying history of old manuscripts. Notice that if we were to cut the edge e' in the tree shown above, the resulting split $A'|B'$, where $A' = \{3, 5\}$ and $B' = \{1, 2, 4, 6, 7\}$, would have the property that $B \cap A'$ is the empty set. This is no accident—it is easy to see that if $A|B$ and $A'|B'$ are arbitrary splits of *any* phylogenetic tree, then one of the four intersections

$$A \cap A', A \cap B', B \cap A', B \cap B'$$

must be the empty set. Buneman showed that this 'pairwise compatibility' condition is not just necessary, but also sufficient for a set of splits to be realized by some phylogenetic tree.

Many splits are well known in biology—for example, we have the vertebrates and invertebrates. We would hope that these splits correspond to splits of the underlying evolutionary tree. More generally, we would like any partition of a set of species to 'fit' their underlying evolutionary tree in the sense that we could delete a set of edges so that the leaves that remain connected form the blocks of the partition. For example, consider the partition $\{\{1, 2\}, \{3, 5\}, \{4, 6, 7\}\}$ and the tree shown above. If we delete both edges e and e', we obtain this partition. We say that the partition is *convex* on the tree.

Convexity has a biological rationale—imagine that the species in each block of the partition have a characteristic (for example, species 1 and 2 have blue eyes, 3 and 5 have red eyes, and 4, 6, and 7 have green eyes). Suppose that this state (eye colour) evolved along the edges of the tree from some hypothetical ancestral vertex in a homoplasy-free way (that is, each state arises just once in the tree). Then the resulting partition will be convex on the tree. Moreover, there is a pleasing converse to this—any partition that is convex on a tree can be

A ***graph*** is a set V of vertices and a set E of 2-element subsets of V. It is much easier to think of a graph as a picture consisting of vertices (points) joined by edges.

A ***path*** in a graph is an alternating sequence of distinct vertices and edges $v_1e_1v_2e_2\cdots e_{k-1}v_k$ so that e_1 joins v_1 and v_2, e_2 joins v_2 and v_3, and so on.

A ***cycle*** is a path in which all vertices are distinct except the first and last which are equal.

A graph is ***connected*** if there is a path joining every pair of vertices.

The ***components*** of a graph are its maximal connected subgraphs.

A ***tree*** is a connected graph with no cycles. Equivalently, a tree is a connected graph that has exactly one more vertex than it has edges.

The ***degree*** of a vertex is the number of edges incident with it.

A ***leaf*** of a tree is a degree-1 vertex.

To ***subdivide*** an edge e of a graph, we replace that edge with a path consisting of two edges. Illustratively, a new vertex is placed at the midpoint of e.

Stirling's formula for approximating factorials is

$$n! \sim \sqrt{2\pi}e^{-n}n^{n+\frac{1}{2}}.$$

A ***partition*** of a set X is a disjoint collection of non-empty subsets of X whose union is X. The non-empty subsets are called *blocks*. A *bipartition* is a partition consisting of two subsets.

described by homoplasy-free evolution on that tree (and from any hypothetical ancestral vertex). Of course, homoplasy can occur for certain characteristics—for example, wings can evolve in some species and then be lost in some later descendant species (reverse evolution), or wings can evolve independently in different lineages (convergent evolution), and so homoplasy-free evolution is an 'ideal' situation. Nevertheless, certain genomic data (such as 'retroposons') exhibit very low levels of homoplasy; this has been helpful in resolving, for example, the tree of placental mammals (including the placement of whales as a sister group to the hippopotamus).

It should be clear that not all sets of partitions can be described by homoplasy-free evolution (i.e. be convex) on the same tree—for bipartitions (splits), the condition for this is precisely Buneman's pairwise compatibility condition described above. For sets of general partitions, the situation is more interesting. Firstly, a pairwise condition won't work—there are sets of partitions where, for each pair of partitions, a tree exists on which both partitions are convex yet for the entire collection, no such tree exists. Indeed, given a set of partitions, deciding whether there exists a tree on which the entire set of partitions is convex is an NP-*complete* problem. Loosely speaking, this means that, in general, the best solution we have is to check each tree T individually to see if each of the partitions in the set are convex on T. Hmmm..., the small needle and large haystack does not make this good news!

Nevertheless, we can ask the following question: For any phylogenetic tree T, what is the smallest collection of partitions required so that T is the only phylogenetic tree on which all the partitions are convex? For this question to make sense, T must be binary (otherwise we could always exhibit another tree that resolves T and that tree would also satisfy the convexity condition for any partition that T does). For splits, the answer to the question is easy—we need one split for each edge of T that is not incident with a leaf (otherwise we could delete that edge and identify its two end-vertices to get a phylogenetic tree on which each split in the collection is still convex). There are $n-3$ such internal edges in a binary phylogenetic tree with n leaves (by Equation (3)), so for splits, we need $n-3$ partitions. The answer to the posed question for sets of general partitions is more interesting. It was recently shown that, for any binary phylogenetic tree T, a set of *at most four* partitions exists so that T is the only phylogenetic tree on which these partitions are convex (i.e. could have evolved without homoplasy). What makes this result surprising is that it does not depend at all on n (the number of leaves of the tree), unlike the $n-3$ result for splits.

Models of Evolution

The result that a tree can be uniquely reconstructed from just four partitions is a combinatorial 'best case' situation. It would be interesting to know how many partitions we would need to reconstruct a tree if these partitions 'evolved' on the tree according to some random model. A particularly simple model is to suppose that each character state (e.g. eye colour) can randomly change on any edge of the tree, and when it does so it always takes on a new state that has not so-far arisen (thereby producing partitions that are convex on the tree).

We can exactly describe the partitions generated in this way by the following simple 'random cluster' process. For each edge e of the tree T, cut edge e with probability p_e or leave it intact with probability $1-p_e$. Perform this process independently across the edges of T and consider the subsets of the leaves of T in the resulting components. This gives us a random partition of the leaves. If we independently generate a large number of partitions in this way, then, with high probability, T will be the only tree on which all the partitions are convex. How many partitions do we need to generate for this to hold? It is unlikely to be a small number (like the 4 we described above) particularly for a large tree, since the partitions are randomly generated, not deliberately chosen by a mathematician! But it still doesn't need to grow too quickly with n (the number of leaves of T); $\log(n)$ turns out to be fast enough, at least under some mild restrictions.

More precisely, suppose that $\varepsilon \leq p_e \leq q$ where $\varepsilon > 0$ and $q < 1/2$ are constants that are independent of n. Then the number of partitions required so that T is (with high probability) the only tree on which all the partitions are convex is:

$$c\frac{\log(n)}{\varepsilon}$$

where c depends just on q and on the probability with which we wish to recover T correctly. When q is greater than 1/2, the number of partitions required for tree reconstruction grows much faster (but still polynomially) with n.

How about models for describing the evolution of individual DNA sites? Since we have just 4 bases, it is unrealistic to expect evolution to be homoplasy-free—if we can change state from (say) A to G, then we should be able later to change state from G back to A. Biologists typically model DNA site evolution as a 4-state Markov process. For example, the tree of the 42 mammals shown above was constructed using such a model in a Bayesian statistical approach.

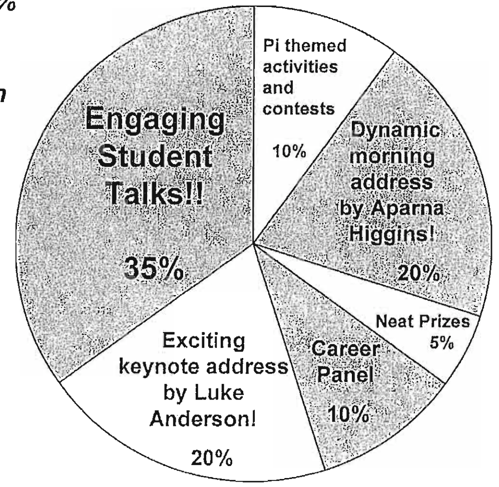
The question of how many sequence sites we need to reconstruct a phylogenetic tree accurately when the sites evolve independently and identically under some finite-state Markov process is difficult. However, it was solved recently by Elchanan Mossel and colleagues at UC Berkeley, and their result is impressive. We saw earlier that a very crude bound (from (7)) requires that the sequence length needed to reconstruct binary phylogentic trees from DNA sequences must grow at the rate $\log(n)$—this argument was based just on counting, and had nothing to do with Markov processes. What Mossel and colleagues have shown is that, under some restrictions on the Markov process, the $\log(n)$ growth rate on sequence length is enough to recover each tree with high probability. In other words, the growth in the sequence length k of at least $\log(n)$ is not only necessary for reconstructing all binary trees on n species (by the crude counting argument), but it is also sufficient, at least for certain Markov processes. The result required developing a fundamentally new approach to tree reconstruction, and an analysis that combined delicate combinatoral and probabilistic arguments. ■

For Further Reading

We have provided just a snapshot of some results in phylogenetics and omitted mentioning many other areas where mathematics plays a crucial role. For more details the reader is referred to *Phylogenetics* (C. Semple and M. Steel, Oxford University Press, 2003). More recent mathematical developments can be found in *Mathematics of Evolution and Phylogeny* (O. Gascuel. (ed.) Oxford University Press, 2005). The last chapter of this book provides the mathematical details described in the final section above. This second book also deals with some complexity of molecular evolution which we have not mentioned; for example, even the use of trees in representing evolution can sometimes be overly simple, and in primitive organisms such as bacteria, processes such as horizontal gene transfer can make a directed acyclic graph a more suitable representation. For a more biological and statistical treatment of phylogenetics, the reader is referred to *Inferring Phylogenies* (J. Felsenstein, Sinauer Press, 2004).

Acknowledgments

We thank Frédéric Delsuc and Nicolas Lartillot for permission to reproduce their mammal tree.


Engaging Student Talks!!
35%
Pi themed activities and contests
10%
Dynamic morning address by Aparna Higgins!
20%
Neat Prizes
5%
Career Panel
10%
Exciting keynote address by Luke Anderson!
20%