Proceedings of Symposia in Applied Mathematics

Phylogenetics

Elizabeth S. Allman and John A. Rhodes

Abstract. Understanding evolutionary relationships between species is a fundamental issue in biology. This article begins with a survey of the many ideas that have been used to construct phylogenetic trees from sequence data. Approaches range from the primarily combinatorial, to probabilistic model-based methods appropriate for developing statistical viewpoints.

The final part of this article discusses a thread of research in which algebraic methods have been adopted to understand some of the probabilistic models used in phylogenetics. Recent progress on understanding the set of possible probability distributions arising from a model as an algebraic variety has helped provide new theoretical results, and may point toward improved approaches to phylogenetic inference.

1. Introduction

Phylogenetics is concerned with inferring evolutionary relationships between organisms. These are depicted by *phylogenetic trees*, or *phylogenies*, whose branching patterns display descent from a common ancestor.

Before the advent of molecular data from biological sequences such as DNA and proteins, construction of a tree for a collection of species required amassing much detailed knowledge of phenotypic differences among them. If fossil evidence of ancestral species was available, it might also be incorporated into the process. Painstaking efforts of experts working for many years were required, yet results might still be controversial, and difficult to justify objectively.

The availability of sequence data produced a revolution in several ways. First, the volume of available data for any given collection of species grew tremendously. Obtaining data became less of a problem than how to sort through it. Second, since sequences are so amenable to mathematical description, it became possible to formalize the inference process, bringing to bear mathematical tools. Although there is still much room for further development of phylogenetics, even a glance at current literature shows that phylogenies inferred from molecular data commonly appear across a large swath of biological fields.

In these notes, we first give a quick survey of the main threads in phylogenetics. As will be apparent, combinatorics, statistics, and computer science have all had

2000 *Mathematics Subject Classification.* Primary 92D15; Secondary 14J99, 60J20.
Key words and phrases. Phylogenetic inference, algebraic statistics, molecular evolution.

large roles to play from the beginning. We conclude with more focused material on recent work in which algebra has provided the framework. We hope that this will provide both an example of how mathematically interesting problems arise in biology, and how various mathematical tools may be brought to bear upon them.

Because of our diverse goals, the level of presentation will vary. We encourage readers to consult the notes on further reading and the bibliography with which we conclude.

The basic problem. Consider the set of species, or *taxa*,

$$X = \{\text{human, chimp, gorilla, orangutan, gibbon}\}$$

that we believe have descended from a common ancestor. If we sequence a gene such as mitochondrial HindIII [**HGH88**] that they all share, we obtain, as the beginning of much longer sequences:

Human	`AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCTCA...`
Chimpanzee	`AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCA...`
Gorilla	`AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCA...`
Orangutan	`AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCC...`
Gibbon	`AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTTCC...`

We have already *aligned* the sequences, so that bases appearing in any column are assumed to have arisen from a common ancestral base. Obtaining a good alignment may be obvious for some datasets, but quite difficult for others, requiring mathematical tools we will not discuss here. We also assume no *deletions* or *insertions* of bases have occurred. In fact, we allow only *base substitutions* where one letter is replaced by another (`A`→`G`, `A`→`C`, etc.)

Similarities in the sequences lend support to our hypothesis of a common ancestor for this gene, while the evolutionary descent has left its record in the differences. Our goal is to pick among all possible phylogenies that might relate these taxa the one that fits 'best' with the data sequences. For instance, two possible trees are shown in Figure 1, and naive consideration of the sequences above might find some support for one over the other.

To be more precise, if X is a set of taxa, a *phylogenetic X-tree* is a tree with its leaves bijectively labeled by elements of X. If an internal node of the tree has been marked to designate the common ancestor, we call that node the *root*, and refer to the tree as a rooted phylogenetic X-tree. Notice that we label only the leaves of the tree, since we generally have no data for any taxa other than those currently living.

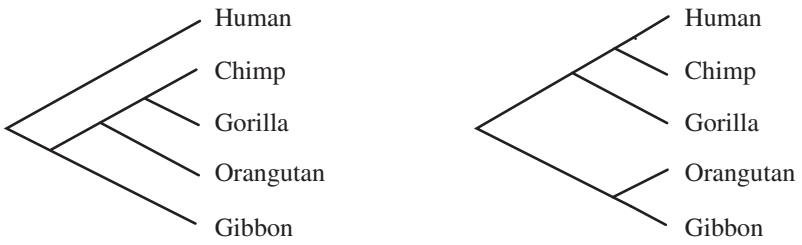

FIGURE 1. Two possible phylogenetic trees.

It is common in biology to focus on binary trees (i.e., trivalent, except bivalent at a root) as being of primary interest. Most speciation events are believed to be of the sort where only two species at a time arise from a parent species. While multifurcations in a tree might be used to represent ignorance (so-called *soft polytomies*), such as when several speciation events occurred so closely in time we are unable to resolve their order, they seldom are believed to represent the true history. For the remainder of this chapter, we consider only binary trees.

The large number of possible trees relating n taxa will turn out to be problematic for most methods of phylogenetic inference. This is quantified in the following basic combinatorial result, easily proved by induction.

THEOREM 1.1. *If* $|X| = n$*, then there are* $(2n-5)!! = 1 \cdot 3 \cdot 5 \cdots (2n-5)$ *distinct unrooted binary phylogenetic* X*-trees, and* $(2n-3)!!$ *distinct rooted ones.*

Before determining a tree that best fits the data we must of course specify what we mean by 'best fits.' There are many approaches to this, and in the next few sections we highlight those that have played the most important roles.

We should add that information in sequences other than base changes can be used to infer phylogenies. Genomes occasionally undergo large scale changes, in which genes may be reordered, duplicated, or lost. Because these changes are rarer, they are all useful for inference much further back in evolutionary time than the base changes we focus on here.

2. Parsimony

One natural criterion for choosing an optimal tree is to find one that requires the fewest base substitutions. The *most parsimonious tree* (or trees) achieves this minimum, and at least in circumstances when substitutions are rare is a reasonable candidate for the best inferred evolutionary history.

Given aligned sequences and any proposed phylogenetic tree relating the taxa, the *Fitch-Hartigan algorithm* can compute the minimal number of substitutions required by that tree. Without proof that the algorithm is correct, we give a brief example, illustrated in Figure 2.

First, the data sequences are placed at the leaves of the tree. We then work upward, filling in possible sequences at adjacent nodes that should attain minimal substitution counts. For instance, at the parent node above the two leaves at the

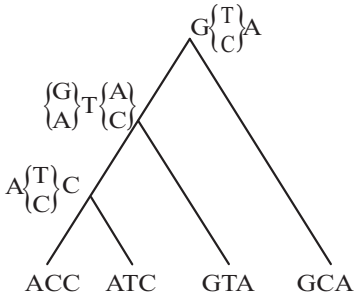

FIGURE 2. The Fitch-Hartigan algorithm for computing parsimony scores.

far left, writing either ATC or ACC would each require only 1 substitution, and we can do no better. We label this node with $\mathtt{A}\{\begin{smallmatrix}\mathtt{T}\\ \mathtt{C}\end{smallmatrix}\}\ \mathtt{C}$, and count that one mutation has occurred. Proceeding to its parent, placing a T in the center site requires no additional substitutions since a T might have occurred in both sequences below it. At the first and third sites, however, substitutions were needed, and all possibilities requiring only 1 substitution per site are listed. So far our substitution count is 3. By filling in sequences at the root, we find we need 1 more substitution, for a total count of 4 for this tree. Thus, 4 is the *parsimony score* of this tree.

The procedure is summarized by: For each node, look at the sequences at its two children. At sites where there are no bases in common, write the union of the sets appearing at the children and increase the substitution count by 1. At sites where there are bases in common, write the intersection of the sets appearing at the children and do not change the substitution count.

Two points should be made that may not be clear from this example: 1) the minimal number of substitutions is independent of root location, so parsimony compares only unrooted trees, and 2) Though it does produce the correct minimal substitution count, the algorithm does *not* reconstruct all ancestral sequences that achieve the minimal count on the tree. Additional steps are needed to do that, if desired.

The Fitch-Hartigan algorithm is fast, in fact $\mathcal{O}(|X|L)$, where L is the number of sites in the sequences. Unfortunately, this is for only one tree, though, and performing it on all trees is more problematic.

THEOREM 2.1 (Foulds and Graham, [**FG82**]). *Determining the most parsimonious tree is NP-hard.*

Branch and bound approaches to searching tree space are sometimes effective, and many heuristics for good searching have been developed and implemented in software. These are believed to perform well in practice, but for a large data set, one never knows for sure that a most parsimonious tree has been found.

A serious problem with parsimony, however, concerns its basic criterion. Suppose that along a single edge of a tree a site evolved as A→C→T. The parsimony criterion would, at best, recognize only one substitution as having occurred. Even worse, for A→C→A it would count no substitutions. If such *hidden mutations* or *back substitutions* occurred, parsimony can be misled.

In fact, using a simple probabilistic model of the substitution process on a small tree (of the sort to be discussed in Section 4) Felsenstein was able to show the following.

THEOREM 2.2 (Felsenstein, [**Fel78**]). *If multiple mutations can occur at a site along any given edge, then there are plausible assumptions under which parsimony will infer the incorrect tree.*

Of course any method of inference may perform poorly when given insufficient data. Felsenstein's result concerns the method's statistical *inconsistency*: Even as the amount of data in accord with the model grows without bound, the wrong tree is inferred.

The inconsistency of parsimony is disturbing to the statistically minded. Nonetheless, parsimony is still in use for inference of trees, though it is not the most popular

method. As long as hidden mutations are believed to be rare, it may be a reasonable approach.

3. Distance methods

The next class of methods share with parsimony a combinatorial flavor. We begin by measuring pairwise *dissimilarity* between taxa, perhaps by using the Hamming distance between their sequences,

$$d(a,b) = \frac{\text{number of sites differing between } a \text{ and } b}{\text{total number of sites}}.$$

We then seek a metric tree, where each edge has a non-negative *length* (or weight), so that cumulative lengths along the tree between taxa (values of the *tree metric*) are close to the dissimilarity values. We view $d(a,b)$ as some sort of measure of how much mutation must have occurred along all edges of the tree between a and b.

Note that we do not refer to the dissimilarity d as a distance or metric, since we should not expect it to agree with a tree metric exactly. Indeed, since we have a finite amount of data (and assuming we believe some stochastic process lies behind it), it will vary from any idealization due to its finiteness and inadequacies of our model. In addition, the Hamming dissimilarity suffers from the same fundamental problem as parsimony — it is insensitive to hidden mutations. After probabilistic models are formalized in Section 4, we will be able to address this last point with improved dissimilarity maps.

Imagine that using the Hamming dissimilarity, or some other measure of difference, we collapse sequence data into a dissimilarity table, like that in Table 1.

Table 1. Dissimilarity between sequences

	a	b	c	d
a		.32	.56	.49
b			.34	.27
c				.37

In algorithmically building a tree from this data, the naive approach is to assume the taxa that are closest in dissimilarity must be closest topologically. This viewpoint applied to Table 1 leads us to join taxa b and d to an ancestral node, and perhaps split their dissimilarity between the two edges. We might then combine b and d into a group, averaging their distances to the other taxa, and thus reduce our table size by one. Continuing in this way, we have outlined the clustering method known formally as UPGMA (Unweighted Pair Group Method using Arithmetic Means).

In fact, UPGMA is *not* a good method in most circumstances in phylogenetics. To see why, note that the tree in Figure 3 *exactly* fits the data of Table 1. However, UPGMA first joins b and d, and so it does not recover this tree. Metric closeness, and topological closeness are in conflict, and so UPGMA made a topological mistake. (UPGMA also produces a rooted tree, with all leaves equidistant from the root. This means all lines of descent from the common ancestor experienced identical amounts of mutation. This implicit *molecular clock* assumption is often not justifiable on biological grounds.)

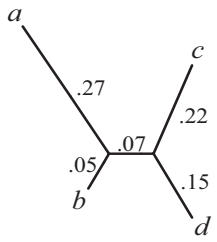

FIGURE 3. A metric tree exactly fitting the dissimilarity data of Table 1.

To address this conflict between metric closeness and topological closeness, the key observation is that for any metric tree with the topology shown in Figure 3, regardless of the edge lengths the following inequality and equality hold:

$$d(a,b) + d(c,d) \leq d(a,c) + d(b,d) = d(a,d) + d(b,c). \tag{3.1}$$

In fact, this leads to a characterization of those dissimilarities that exactly fit metric trees.

THEOREM 3.1 (Buneman, [**Bun71**]). *A dissimilarity map d on X arises from a metric tree if, and only if, for every choice of 4 taxa $a, b, c, d \in X$, the following 4-point condition holds:*

$$d(a,b) + d(c,d) \leq \max(d(a,c) + d(b,d), d(a,d) + d(b,c)).$$

EXERCISE 3.2. Prove Theorem 3.1 in the special case of 4-leaf trees.

The *Neighbor Joining algorithm* of Saitou and Nei [**SN87, SK88**] overcomes the failings of UPGMA by efficiently finding a pair of taxa a, b to join that would, for data exactly fitting a tree, satisfy Equation (3.1) for all pairs c, d. It thus is guaranteed to produce the correct tree for 'perfect' data, and has been found to perform well on simulated data. With running time $\mathcal{O}(|X|^3)$ when dissimilarities are already computed, it is quick, since it need not search among all possible trees. NJ is widely used when a tree must be produced quickly, and is by far the most popular distance-based method.

One criticism of NJ is that while its algorithmic approach is fast, it is unclear what it optimizes: In what sense do we get the best tree? Although there are other distance approaches with explicit optimality criteria (e.g., best L^2 fit, best L^1 fit, a 'minimum evolution' criterion), implementations in software then require searching tree space, so the speed advantage of NJ is lost.

Another issue is that if a distance approach only compares sequences two-at-a-time, it is ignoring much of the information in the data. Recent work [**PS04a, CL05**] has sought to overcome this issue partially, yet still preserve some of the speed of NJ, by considering sequences k-at-a-time.

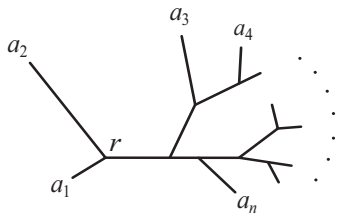

FIGURE 4. An n-taxon tree.

4. Base Substitution Models

Before going further with our survey of common phylogenetic methods, we must introduce some of the probabilistic models of molecular evolution which other methods use. Explicit models allow a firmer grounding in statistical theory.

Most probabilistic models of the mutation process focus on a single site in a sequence, and only on base substitutions occurring at that site as evolution proceeds down a tree. Other types of sequence changes — insertions, deletions, inversions — require more complicated models than will be discussed here.

To introduce the form of the model, consider some fixed rooted tree such as the one in Figure 4. At the root node, our site might have any of the 4 bases A, G, C, T occurring. A *root distribution vector* $\boldsymbol{\pi}_r = (\pi_A\ \pi_G\ \pi_C\ \pi_T)$ gives the probabilities of each occurring. On an edge e leading from the root, substitutions may occur, so a 4×4 Markov matrix M_e specifies the 16 conditional probabilities of the various substitutions $\mathtt{A}\to\mathtt{A}$, $\mathtt{A}\to\mathtt{G}$, etc. From $\boldsymbol{\pi}_r$ and M_e we can find the probabilities of the various bases at the descendent node at the end of e. Thus if we specify a Markov matrix for each edge of the tree, we have modeled how the entire evolutionary process proceeds over the tree.

In formalizing this we model sequences built of an arbitrary κ-letter alphabet. For each node of the tree we have a random variable which might assume any of κ states, usually denoted by the elements of $[\kappa] = \{1, 2, \ldots, \kappa\}$. The root distribution vector $\boldsymbol{\pi}_r$ gives probabilities of the various states for the variable at the root, while $\kappa \times \kappa$ Markov matrices give transition probabilities of state changes from ancestral to descendent node along each edge. Since an n-leaf trivalent tree has $2n-3$ edges, this number of Markov matrices must be specified. The parameters for the *general Markov model* (GM), are then

(1) a leaf-labeled tree T,
(2) a root distribution vector $\boldsymbol{\pi}_r$ with non-negative entries summing to 1, and
(3) a Markov matrix M_e (non-negative entries, each row summing to 1) for each edge e.

For DNA, the number of states is $\kappa = 4$, but for protein sequences, which are built from twenty amino acids, $\kappa = 20$. The case $\kappa = 2$ is also of interest for DNA substitution models, if we group bases into *purines* $R = \{A, G\}$ and *pyrimidines* $Y = \{C, T\}$. We often refer to $(\boldsymbol{\pi}, \{M_e\})$ as the *stochastic parameters*, distinguishing them from the tree parameter.

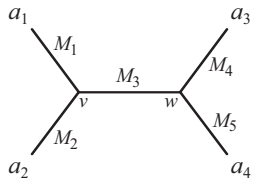

FIGURE 5. Computing the expected pattern frequencies on T.

A key point in the use of a model such as this is that while it describes states at all nodes of the tree, in fact only those at the leaves are observable, since the leaves represent the extant taxa from which we may obtain data.

With the parameters of the model thus specified, we are interested in the *joint distribution P of states at the leaves* $a_1, a_2, \ldots, a_n$. The joint distribution P is an n-dimensional $\kappa \times \cdots \times \kappa$ tensor (or table or array) with entries

$$P(i_1, \cdots, i_n) = \text{Prob}(a_1 = i_i, \cdots, a_n = i_n).$$

The entries of P then are the expected frequencies of observing a *pattern* of states such as $(i_1, \cdots, i_n)$ at the leaves of the tree. These *expected pattern frequencies* can be explicitly expressed in terms of the parameters of the model, as we explain through an example.

EXAMPLE 4.1. Consider the 4-taxon tree of Figure 5 rooted at v, with stochastic parameters as labeled. Using α and β to represent the unobserved states at the two internal nodes v and w, respectively, the expected pattern frequency $P(i, j, k, l) = p_{ijkl}$ is given by

$$p_{ijkl} = \sum_{\beta=1}^{\kappa} \sum_{\alpha=1}^{\kappa} \pi_\alpha M_1(\alpha, i) M_2(\alpha, j) M_3(\alpha, \beta) M_4(\beta, k) M_5(\beta, l).$$

Note the form of this expression depends very much on the topology of the tree, and in fact the topology can be recovered from the formula.

While the model outlined here describes the base substitution process at a single site, for phylogenetic inference the data is aligned DNA sequences of some length L. To apply the model to data, we make the additional assumption that each site in the aligned sequences is a trial of the same probabilistic process. More carefully, we assume that the evolutionary process at each site proceeds independently of all other sites, but according to the same probabilistic process, with the same parameters.

This independent, identically distributed (i.i.d.) assumption is not desirable from a biological viewpoint — substitutions at one site may well not be independent of nearby sites, or even of distant sites if the three-dimensional structure of a protein coded for by the gene folds to bring distant stretches together. Also, allowing different substitution processes might better describe what goes on in the evolution of various parts of the sequence.

Nonetheless, some form of the i.i.d. assumption is essential. It is only by viewing each site as a trial of the same process that we obtain enough data to infer something about the parameters. With this assumption, we can estimate the expected pattern frequencies such as p_{ijkl} by the observed frequencies of patterns in the aligned sequences $\widehat{p}_{ijkl}$. Then formulas such as that in Example 4.1 play a role in our

FIGURE 6. A 2-taxon tree.

inference of the root distribution, Markov matrices, and most importantly, the tree.

We now wish to show that for most parameter choices we can produce the same joint distribution at the leaves of a tree as we could with a different root location and a related choice of parameters.

To develop this idea, first consider the 2-taxon tree of Figure 6, with a_1 designated as the root. Let $\boldsymbol{\pi}_{a_1} = (\pi_1\ \pi_2\ \pi_3\ \pi_4)$ be the root distribution vector, and, for $e = (a_1 \to a_2)$, let

$$M_e = (m_{ij}), \quad m_{ij} = \text{Prob}(a_2 = j \mid a_1 = i),$$

be the matrix of conditional probabilities of base substitutions along the edge.

To compute the joint distribution $P = P_{a_1a_2}$, a 4×4 matrix of expected pattern frequencies, notice that the (i,j)-entry p_{ij} is $\pi_i m_{ij}$, or in matrix form,

$$P_{a_1a_2} = \text{diag}(\boldsymbol{\pi}_{a_1})M_e = \begin{pmatrix} \pi_1m_{11} & \pi_1m_{12} & \pi_1m_{13} & \pi_1m_{14} \\ \pi_2m_{21} & \pi_2m_{22} & \pi_2m_{23} & \pi_2m_{24} \\ \pi_3m_{31} & \pi_3m_{32} & \pi_3m_{33} & \pi_3m_{34} \\ \pi_4m_{41} & \pi_4m_{42} & \pi_4m_{43} & \pi_4m_{44} \end{pmatrix}, \tag{4.1}$$

where $\text{diag}(\boldsymbol{\pi}_{a_1})$ denotes the diagonal matrix with entries from $\boldsymbol{\pi}_{a_1}$.

Now consider the same 2-taxon tree T in Figure 6, but with the root taken at a_2 instead. Then in terms of the stochastic parameters on T rooted at a_1, define the root distribution vector to be $\boldsymbol{\pi}_{a_2} = \boldsymbol{\pi}_{a_1}M_e$, the probabilities that leaf a_2 is in each of the four states. Let $M_{e'}$ denote a Markov transition matrix for $e' = (a_2 \to a_1)$ that will be determined shortly.

Notice that viewing a_2 as the root, the joint distribution is expressed as $P_{a_2a_1} = (P_{a_1a_2})^T$. Thus we would like to find $M_{e'}$ so that,

$$\text{diag}(\boldsymbol{\pi}_{a_2})M_{e'} = P_{a_2a_1} = (P_{a_1a_2})^T = (\text{diag}(\boldsymbol{\pi}_{a_1})M_e)^T = M_e^T\,\text{diag}(\boldsymbol{\pi}_{a_1}).$$

If the entries of $\boldsymbol{\pi}_{a_2}$ are all positive, then we may take

$$M_{e'} = \text{diag}(\boldsymbol{\pi}_{a_2})^{-1}M_e^T\,\text{diag}(\boldsymbol{\pi}_{a_1}).$$

This establishes that, under mild conditions, there is a choice of parameters for T rooted at a_2 that give rise to the same joint distribution P as the parameters $\boldsymbol{\pi}_{a_1}$ and M_e for T rooted at a_1. Hence, for the general Markov model, we can 'move the root' without affecting the entries of the joint distribution array P. We formalize these observations and extend the setting to n-taxon trees in Proposition 4.2.

PROPOSITION 4.2. *Fix an n-taxon tree T. Let r be some choice of root for T (which may be a leaf, an internal node of valance 3, or along some edge). Then, for generic choices of stochastic parameters S_r for the general Markov model rooted at r, and for any other choice of a root s for T at either a leaf or an internal node of valance 3, there is a uniquely determined choice of general Markov model*

parameters S_s for the model rooted at s producing the same joint distribution at the leaves as S_r.

A consequence of Proposition 4.2 is that the location of the root in a tree T is a *biological* problem, not a mathematical one. Under this model (and many others as well), there is no way to mathematically identify a node in T as a most recent common ancestor of the taxa in hand. (However by including an *outgroup*, a taxon known to be distantly related to those under study, one can use biological knowledge to locate a root.)

For this reason, we usually consider the inference of an unrooted tree as our goal. In addition, for computations with a model, and in arguments, we are now free to place roots wherever we find most convenient.

While the general Markov model is simple to explain, it has more parameters than models typically used in practice. Once the tree parameter has been chosen as a particular n-taxon tree, there are $\kappa - 1$ free choices to be made for $\boldsymbol{\pi}_r$, and for each of the $2n-3$ edges, $\kappa(\kappa-1)$ free choices for entries in M_e, giving a total of $\kappa - 1 + (2n-3)\kappa(\kappa-1)$ numerical parameters. Though this grows only linearly in the number of taxa, the coefficient is rather large. For $\kappa = 4$, the total number of parameters is already roughly $24n$.

This large number of parameters has two effects. First, it slows down computations, which for a large number of sequences can be problematic. Second, using a parameter-rich model allows us to better fit data, but may also allow overfitting. If the data can be described by a model with fewer parameters, that model may provide a better basis for inference.

Restrictions on the particular form of the stochastic parameters, some arising from biological considerations and some for mathematical convenience, give rise to submodels of the GM model. We discuss these, as well as some extensions to more elaborate models next.

Group-based models. The *Jukes-Cantor* model for DNA is the biologically-plausible model with the fewest parameters. It assumes a uniform root distribution vector of $\boldsymbol{\pi} = (.25\ .25\ .25\ .25)$ and edge transition matrices of the form

$$M_{JC} = \begin{pmatrix} 1-a & \frac{a}{3} & \frac{a}{3} & \frac{a}{3} \\ \frac{a}{3} & 1-a & \frac{a}{3} & \frac{a}{3} \\ \frac{a}{3} & \frac{a}{3} & 1-a & \frac{a}{3} \\ \frac{a}{3} & \frac{a}{3} & \frac{a}{3} & 1-a \end{pmatrix},$$

where a different value of a may be used for each edge. On each edge, all non-identical base substitutions are equally likely, and the probability that some change occurs at a site between the endpoints of the edge is given by the parameter a. Note that the root distribution is an eigenvector of M_{JC}, so a uniform distribution of states will occur at each node in the tree.

Though this model is attractive for its simplicity, further realism can be introduced by having two probabilities of changes, as we now describe. Because of chemical similarities, the bases are classified as purines $\{A, G\}$ and pyrimidines $\{C, T\}$. Assigning probability a to in-class changes (transitions), and b to out-of-class changes (transversions), we arrive at the *Kimura 2-parameter* model, with

matrices

$$M_{K2P} = \begin{pmatrix} 1-(a+2b) & a & b & b \\ a & 1-(a+2b) & b & b \\ b & b & 1-(a+2b) & a \\ b & b & a & 1-(a+2b) \end{pmatrix},$$

where the rows and columns are ordered by the states A, G, C, T, (purines, followed by pyrimidines). Typically $a > b$, since transitions are often observed more frequently than transversions.

A slight generalization, introduced more for its mathematical structure than for biological reasons, is the *Kimura* 3*-parameter* model with transition matrices of the form

$$M_{K3P} = \begin{pmatrix} 1-(a+b+c) & a & b & c \\ a & 1-(a+b+c) & c & b \\ b & c & 1-(a+b+c) & a \\ c & b & a & 1-(a+b+c) \end{pmatrix}.$$

Notice the pattern to the entries is that of an addition table for the group $\mathbb{Z}_2 \times \mathbb{Z}_2$. In fact, if we identify the four bases with group elements by

$$A = (0,0),\ G = (1,0),\ C = (0,1),\ T = (1,1),$$

then a substitution $X \to Y$ is naturally encoded by the group element $Y - X$ since $X + (Y - X) = Y$. If a random choice of a group element determines what substitution occurs, then the numbers $1-(a+b+c)$, a, b, and c represent the probabilities of each choice, explaining the form of M_{K3P}.

This special structure has produced some quite interesting results for the K3P model, and its specializations, the K2P and JC models. Although we will not explain it here, the *Hadamard conjugation* of [**Hen89, HP89**] is a fundamental result that introduced Fourier analysis as a tool for studying such models. This was further developed in [**SSE93**].

General Time Reversible models. So far we've essentially taken a discrete time approach to modeling substitutions, by specifying transition probabilities relating states at the two ends of an edge. Substitutions may also be modeled as a continuous time process. In fact, if we are interested in inferring elapsed time between speciation events, we must take this approach so that those times become parameters.

To formulate a continuous time model, we let Q denote a $\kappa \times \kappa$ instantaneous *rate matrix*. The off-diagonal entries of Q represent rates at which the 12 non-identical substitutions occur, and are thus non-negative numbers. The diagonal entries are chosen so that the rows sum to zero. Associated to each edge e of T is a parameter t_e, an edge length. If $e = (v \to w)$, then t_e represents the amount of time elapsed during evolution of the sequence at v into the sequence at w. The Markov transition matrix for e is then the matrix exponential, $M_e = \exp(Qt_e)$.

When using continuous time models, we generally choose one rate matrix Q for all edges of the tree. This imposes some commonality to the evolutionary process on all edges of the tree that is biologically reasonable in some (but not all) circumstances. It also dramatically reduces the dependency of the number of parameters of the model on the number n of taxa to roughly $2n$, since that is the

growth rate of the number of edges, and we add only one new parameter, t_e, for each edge.

It's usually most convenient to require that the root distribution vector $\boldsymbol{\pi} = \boldsymbol{\pi}_r$ be an eigenvector of Q with eigenvalue 0. This ensures that $\boldsymbol{\pi}$ is an eigenvector of $M_e = \exp(Qt_e)$ with eigenvalue 1, for all values of t_e. As a result, the model is a *stationary* one, with state distribution the same at all nodes of the tree.

Along with this we often assume *time-reversibility*:

$$\operatorname{diag}(\boldsymbol{\pi})Q = Q^T \operatorname{diag}(\boldsymbol{\pi}).$$

Imposing this condition, that $\operatorname{diag}(\boldsymbol{\pi})Q$ is symmetric, implies that $\operatorname{diag}(\boldsymbol{\pi}) \exp(Qt_e)$ is symmetric for all t_e. But we saw in Equation (4.1) that $\operatorname{diag}(\boldsymbol{\pi}) \exp(Qt_e)$ represents the joint distribution of states at the two ends of e. Therefore, stationarity with time-reversibility means that we can use the same parameters $\boldsymbol{\pi}, Q, t_e$ to model evolution on an edge regardless of the orientation of the edge.

The *general time-reversible* (GTR) model is a rate-matrix model making both the stationarity and time-reversible assumptions. These assumptions imply we can 'move the root' in a tree under the GTR model without affecting the entries of the joint distribution P. As with the GM model, we will not be able to mathematically determine a root location when we use the GTR model for inference. This is convenient, since it means for inference we will not have to search over all rooted trees, but rather over unrooted ones. In fact, the reason the GTR model is used is precisely that it is the most general rate-matrix model with this property.

EXERCISE 4.3. Show that all pairs $\boldsymbol{\pi}, Q$ for the κ-state GTR model can be specified by formulas involving $(\kappa - 1) + \kappa(\kappa - 1)/2$ scalar parameters, and thus, after normalizing so that one edge has length 1, the GTR model on an n-taxon tree has $(\kappa - 1) + \kappa(\kappa - 1)/2 + (2n - 4)$ parameters.

EXERCISE 4.4. Show that the JC model is a special case of GTR, by finding $\boldsymbol{\pi}$, Q_{JC} explicitly.

EXERCISE 4.5. Show that the K2P model may be a special case of GTR, but that there are choices of K2P parameters that are not instances of a GTR model. More specifically, find $\boldsymbol{\pi}$ and all possible Q so that $M_e = \exp(Qt_e)$ are K2P edge transition matrices. Then find a relationship between the sets of eigenvalues of M_e, for each of the $2n - 3$ edges e, that must hold if an arbitrary K2P model is a GTR model.

EXERCISE 4.6. Consider the 2-taxon tree of Figure 6 rooted at a_1. If $\boldsymbol{\pi}_{a_1} = (.23\ .26\ .24\ .27)$ and $M_{a_1a_2} = \begin{pmatrix} .96 & .02 & .01 & .01 \\ .06 & .88 & .02 & .04 \\ .01 & .04 & .93 & .02 \\ .05 & .05 & .04 & .86 \end{pmatrix}$, compute the joint distribution of bases at the leaves $P = P_{a_1a_2}$. Could this come from the GTR model?

Mixture models. It is unrealistic biologically to assume that all sites mutate according to the same process. Certainly it is plausible that non-coding regions of the genome might undergo substitutions at a faster rate than coding ones. But even within genes, there may be variability.

Triplets of bases form *codons* that specify an amino acid to appear in the protein molecule for which the gene encodes, but the *genetic code* which relates codons to amino acids has redundancy. There are 4^3 possible codons, but only 20 amino acids. Much of the redundancy of the code is such that different bases in the third codon position may not affect the gene product. Thus the third position may be more likely to experience a higher mutation rate.

Moreover, since some parts of the protein structures might be essential to the viability of an organism, those sites coding for such parts may never be observed to undergo any substitutions at all. Typically, then, we expect variability in the mutation process among sites, but do not know how to partition the sites into various classes according to their behavior.

A step toward improving our description of molecular evolution then is to introduce a *mixture model*. In this formulation, each site in aligned sequences falls into one of k classes and each of the k classes carries its own stochastic parameters. Of course all sites share the same tree parameter. An additional $k-1$ parameters δ_i are needed to indicate the proportion of the sites that lie in the ith class, $i = 1, \cdots, k-1$, with $\delta_k = 1 - \sum_{i=1}^{k-1} \delta_i$ giving the proportion in the last class.

For example, consider a situation in which some sites in our sequences are believed to be unable to undergo substitutions, perhaps because of functional constraints on a protein product. We call these *invariable* sites. Notice that in aligning sequences we usually have many sites that are in complete agreement, but we do not necessarily believe they were invariable — they may have been able to undergo substitutions, but simply did not. If we believe invariable sites exist, then we cannot directly distinguish between the constant sites which are invariable and those which were free to vary but did not.

To model this, we introduce the GM+I model. For a fixed rooted tree, the parameters are 1) For the sites that can vary, a root distribution vector $\boldsymbol{\pi}_{GM}$ and Markov matrices $\{M_e\}$ for each edge of the tree, 2) For those sites that are invariable, a root distribution vector $\boldsymbol{\pi}_I$, and 3) a mixing parameter δ indicating the proportion of sites that mutate according to a GM process. The resulting joint distribution P is a weighted sum

$$P_{GM+I} = \delta P_{GM} + (1-\delta) P_I,$$

where P_{GM} is the joint distribution for the varying sites, and P_I the joint distribution for the invariable sites, an n-dimensional diagonal tensor formed from $\boldsymbol{\pi}_I$.

GM+I is just one of many mixture models that can incorporate more biological realism in modeling the base substitution process, and this example should make clear what we mean by mixture models such as GM+GM+GM, GM+GM+I, GTR+I, or JC+I.

The use of mixture models can greatly increase the number of stochastic parameters. For a k-class general Markov mixture model, for instance, the number of stochastic parameters increases by more than a factor of k. While mixture models are appealing since they conform better with our intuitive notion of how to model base substitutions, the large number of parameters increases the risk of overfitting data. At the extreme, one might imagine a mixture model with so many classes that, for appropriate parameter choices, it might be capable of producing *any* joint distribution at all. If such a model describes data, then we have no hope of inferring

a tree, since no signal indicating the correct tree can be found in the observed joint distribution.

One model of substitutions that is in widespread use, the GTR+I+Γ, is a more restricted version of the mixtures above. It is a *rates-across-sites model*, and cuts something of a compromise between mixing classes and keeping the number of parameters down. Here a root distribution $\boldsymbol{\pi}$, an instantaneous rate matrix Q, edge lengths t_e, and a mixing parameter for the variable and invariable classes are specified as in a GTR+I model. In addition, a Γ distribution describes the distribution of a *rate parameter* λ for the different variable sites, with each site undergoing substitution along an edge e according to $M_e = \exp(Q\lambda t_e)$. Thus while we introduce a continuum of variable classes, we need only one new parameter, the shape parameter for the Γ distribution. The model therefore assumes much commonality to the substitution process, since the same Q is used on all edges and for all variable sites. In practice, when a model such as GTR+I+Γ is used for inference, it must be incorporated into software, and that means the Γ distribution is discretized and only a small number of discrete classes are used.

Note that we have given no biological justification for preferring the Γ distribution to any other distribution. Indeed, there seems to be none. It is simply hoped that by tuning the shape parameter, the distribution is flexible enough to capture whatever variation in rates might exist.

5. Improved Dissimilarities from Models

With a probabilistic model in hand, we can sometimes create *phylogenetic distances* that better measure the amount of mutation that occurred in the evolution of two sequences from their common ancestor. These can be used in place of the Hamming dissimilarity in distance methods of inference, such as Neighbor Joining.

We sketch the idea for the Jukes-Cantor model with $\kappa = 4$. Suppose the evolution of an ancestral sequence for taxon a_1 to a descendent sequence for taxon a_2 is modeled by the Jukes-Cantor model. Then we begin with a uniform distribution $\boldsymbol{\pi} = (.25\ .25\ .25\ .25)$ of states for a_1, and substitution occurs according to a Markov matrix of the form

$$M_{JC} = \begin{pmatrix} 1-a & \frac{a}{3} & \frac{a}{3} & \frac{a}{3} \\ \frac{a}{3} & 1-a & \frac{a}{3} & \frac{a}{3} \\ \frac{a}{3} & \frac{a}{3} & 1-a & \frac{a}{3} \\ \frac{a}{3} & \frac{a}{3} & \frac{a}{3} & 1-a \end{pmatrix} = \exp(Qt), \tag{5.1}$$

where t represents the amount of time of evolution along e and Q is the rate matrix

$$Q = \begin{pmatrix} -1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -1 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & -1 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -1 \end{pmatrix}.$$

By our choice of Q, we have chosen to measure time in such a way that the instantaneous rate at which (non-identical) base-substitutions occur is 1.

Exercise 5.1. By diagonalizing, show that Equation (5.1) implies

$$t = -\frac{3}{4}\ln\left(1 - \frac{4}{3}a\right).$$

Since a represents the probability of a (non-identical) substitution being observed when we compare a site in the sequences for a_1 and a_2, we can estimate a by the Hamming distance between the sequences, $\widehat{a}$. We thus define the *Jukes-Cantor distance* between the sequences as

$$d_{JC}(a_1, a_2) = -\frac{3}{4} \ln \left(1 - \frac{4}{3}\widehat{a}\right).$$

This distance (which is really a dissimilarity, and not likely to be exactly in accord with a tree metric when computed from data) is a measure of the total amount of mutation per site that occurred between a_1 and a_2, including all those unobserved substitutions which were hidden by subsequent substitutions. Its value is larger than $\widehat{a}$ to account for these.

Notice we assumed one of our sequences is the ancestral one, and when dealing with data we typically have sequences only from extant species. However, the Jukes-Cantor model is in fact a special case of the GTR model, so we may freely choose any node on a tree as the root. With a little thought, we see that d_{JC} calculates the total substitutions per site that occurred in both lineages descending to a_1 and a_2 from their common ancestor.

A more complete derivation of the Jukes-Cantor distance would show that it is the *maximum likelihood* estimate for the total amount of substitution that occurred between two sequences under the JC model. This is important, since it means its use for estimating amounts of mutation is well-founded in statistical theory.

By similar means, phylogenetic distances can be defined for the Kimura models and others. In practice, these are the sorts of distances that are used when a method such as Neighbor Joining is used to construct a tree. Since d_{JC} is estimating the cumulative edge lengths between taxa, it should give distances that are close to exactly fitting a tree, at least to the extent that the Jukes-Cantor model describes our data accurately, and our sequences are sufficiently long so that $\widehat{a} \approx a$.

However, distance formulas have not been discovered for all models. For instance, for a GTR model with a rate distribution, such as GTR+Γ, a distance formula can be given once one has specified the distribution. However, no distance is known that is appropriate for an unknown distribution. Other mixture models, such as GTR+I, also lack known distance formulas.

Use of an appropriate distance can improve a tree construction method such as Neighbor Joining, making it statistically consistent. However, it does nothing to address the problem that distances are based on two-sequence comparisons, and therefore do not make full use of the data.

6. Statistical inference via Maximum Likelihood

With a model of mutation specified, we can also infer trees from data by using the maximum likelihood approach not just to find distances between two taxa at a time, but rather to estimate all the parameters of the model. The tree parameter is of course the one we are most interested in, but we can not separate it from the others in this approach.

For a fixed model (say GTR+I+Γ) a full maximum likelihood estimation of all parameters from aligned sequences for n taxa would proceed as follows:

For each pattern $\mathbf{i} \in [\kappa]^n$, let $n_{\mathbf{i}}$ denote the number of sites in the aligned sequences in which that pattern is observed. Then

(1) Loop on all $(2n-5)!!$ unrooted phylogenetic trees that might relate the taxa.
(2) For each tree T, viewing the joint distribution at the leaves, P, as a function of the stochastic parameters, consider the *likelihood function*

$$\begin{aligned} L(T, &\text{stochastic parameters}) \\ &= \text{Prob}(\text{data} \mid T, \text{stochastic parameters}) \\ &= \prod_{\mathbf{i}\in[\kappa]^n} P_{\mathbf{i}}^{n_{\mathbf{i}}}. \end{aligned}$$

Determine the maximum value of this function, which is the likelihood of the tree T.
(3) Report as the maximum likelihood tree the T which has the greatest likelihood.

Obviously this scheme cannot be carried out exactly if the number of taxa is large. First, there are too many trees to consider each, and so heuristic searches among the trees must be performed. Second, even computing the likelihood for one tree is difficult, since we must solve a multivariate optimization problem. This is one reason why keeping the number of parameters in the model small is so desirable. We must also be aware that attempts to optimize the numerical parameters may find only local maxima, and fail to find the true global maximum.

Nonetheless, computer implementations of maximum likelihood inference are heavily used because of the desirable statistical properties of the method. But as the number of taxa is increased, it becomes impossible to complete the searches in a reasonable amount of time.

Recently there has also been rapidly growing interest in using Bayesian approaches to phylogenetic inference as well, and software is available built on MCMC algorithms. Since a complete survey should certainly outline this approach, we suggest [**Gas05**] for a good overview.

7. Algebraic Methods in Phylogenetics

Although phylogenetic inference is regularly conducted by the methods outlined above, there is still much potential to improve both our methods and understanding of the problem. Even if one adopts a preferred inference method, be it parsimony, maximum likelihood, or Bayesian approaches, the computational issues in performing the method force compromises in carrying out the procedure. Any new perspectives that can be developed have potential to guide us toward better approaches.

In recent years, the perspectives of algebraic geometry have been brought into phylogenetics. Though still very much under development, we turn now to introducing this viewpoint.

Phylogenetic invariants, ideals, and varieties. Consider a fixed n-taxon tree T. Then any probabilistic model of molecular evolution on T defines a map from stochastic parameter space to joint distribution space. For the GM model on an n-taxon tree, for example, the number of stochastic parameters is $K = (\kappa-1)+(2n-3)\kappa(\kappa-1)$, so the stochastic parameter space S is a subset of $[0,1]^K$.

Since the joint distribution P is an n-dimensional array, the GM model on T defines a map ϕ_T:

$$\begin{aligned} \phi_T : S &\longrightarrow [0,1]^{\kappa^n} \\ (\boldsymbol{\pi}, \{M_e\}) &\longmapsto P. \end{aligned}$$

EXAMPLE 7.1. Recall Example 4.1, for a 4-taxon tree. There we saw that the map ϕ_T was defined by the component functions

$$P(i,j,k,l) = \sum_{\beta=1}^{\kappa} \sum_{\alpha=1}^{\kappa} \pi_\alpha M_1(\alpha,i) M_2(\alpha,j) M_3(\alpha,\beta) M_4(\beta,k) M_5(\beta,l),$$

so that each component function is a degree 6 polynomial in the scalar parameters, with κ^2 terms.

More generally, for the GM model on an n-taxon tree each of the component functions of ϕ_T will be a degree $2n-2$ polynomial, with κ^{n-2} terms. The precise form of these polynomials reflects the topology of the tree T.

The fact that the function ϕ_T is polynomial suggests extending it beyond the stochastic setting, to the complex numbers. Accordingly, if $S \subset \mathbb{C}^K$, then we have a complex parameterization map

$$\begin{aligned} \phi_T : \mathbb{C}^K &\longrightarrow \mathbb{C}^{\kappa^n}, \\ (\boldsymbol{\pi}, \{M_e\}) &\longmapsto P, \end{aligned}$$

defined by the same polynomial formulas. Here we are simply allowing $\boldsymbol{\pi}$ and M_e to have complex entries.

DEFINITION 7.2. The *phylogenetic variety* for the GM model on T is $V_T = \overline{\mathrm{Im}(\phi_T)}$, where the bar denotes (Zariski and standard) closure.

DEFINITION 7.3. For any phylogenetic variety V_T, let I_T be the ideal of all elements of the polynomial ring $\mathbb{C}[P]$ in κ^n variables that vanish on V_T. Then I_T is the *phylogenetic ideal*, and its elements are called *phylogenetic invariants*.

In essence, the phylogenetic variety V_T is a higher dimensional 'surface' that contains the (complex) joint distribution for all possible choices of (complex) numerical parameters $s = (\boldsymbol{\pi}, \{M_e\})$ of GM on T.

One original motivation for studying phylogenetic varieties is that they group together into one object, V_T, all joint distributions for a model that are associated to a particular tree topology. In applications, the tree topology is usually the parameter of greatest interest. If an observed distribution of pattern frequencies were 'close' to V_T, that could be interpreted as support for inferring T. The vanishing, or rather near-vanishing, of phylogenetic invariants could indicate 'closeness,' thus potentially allowing the inference of T without having to estimate *all* the other parameters, as maximum likelihood requires. This would decouple the tree inference problem from the problem of inferring all numerical parameters.

This approach to inference is still largely unrealized, however. It will, at the very least, require considerable more sophistication than it has been presented with here. If we want to check for the 'near vanishing' of invariants, and invariants form an ideal, we might first consider only a finite set of invariants forming a basis for the ideal. But what basis should we choose? It is not clear what a good choice would be, but what choice we make will of course be reflected in the values the

polynomials take on on data. Naive approaches to judging 'near vanishing' for one basis may not correspond to 'near vanishing' for another. And regardless, all of this should be grounded in some sort of statistical reasoning so we can understand better how it should perform with data.

REMARK 7.4. Extending the parameterization ϕ_T to the complex numbers from the stochastic setting is done because an algebraically closed field provides the easiest and most natural setting for understanding polynomial maps. Of course complex parameters and complex joint distributions P are not so natural from a biological or statistical viewpoint. As the goal ultimately is to understand the GM model in a stochastic setting, a more appropriate setting might be real algebraic geometry. That study, however, remains for the future.

Also, by taking the closure, some points in V_T have been introduced that are not in the image of the parameterization ϕ_T. While this is natural to an algebraic geometer, we can also justify it in another way. If a point is in this closure, then there are points on the parameterized portion of V_T that are arbitrarily close to it. If we have an observed joint distribution from data sequences, and we are trying to determine if it is 'close' to the variety, then it makes no difference whether we see if it is 'close' to the parameterized portion of the variety, or to any point on the variety.

One can of course define phylogenetic varieties and invariants for other models, such as JC or K3P, which have polynomial parameterization maps as well. For most models, these varieties are irreducible (equivalently, the phylogenetic ideal is prime), but see [**AR06**] for an exception. We continue to focus on the GM model for most of our exposition.

Phylogenetic invariants were introduced independently in two papers, by Cavender and Felsenstein [**CF87**], and by Lake [**Lak87**], both for simpler models than GM. Lake dealt only with linear invariants, while Cavender and Felsenstein considered higher degree ones as well, and even dealt with some issues of real vs. complex geometry. Though using no language of algebraic geometry, [**CF87**] is still an excellent introduction to the viewpoint.

Finding invariants. The dimension of stochastic parameter space, K, is much smaller than the dimension of joint distribution space κ^n, and as a result there should be many polynomials vanishing on V_T. How to find them explicitly, though, is not obvious.

But finding phylogenetic invariants is simply an instance of an implicitization problem in algebraic geometry: Given a parameterized variety such as V_T, with a polynomial parameterization map ϕ_T, find an implicit description of it as the zero set of polynomials. Once we have fixed a choice of model and T, we can write down explicit formulas for the map ϕ_T. Then implicitization can be attempted computationally, as a variable elimination problem using Gröbner bases (see, for instance, [**CLO97**]).

As long as the model is simple (a small number κ of states and a small number of parameters), and the tree is small (so the dimension κ^n of the space in which V_T lies is not too large), this can be done by software such as Maple, Macaulay2 [**GS02**], Singular [**GPS01**], or other computational algebra packages. However, one quickly reaches the limits of current software as the number of states, the

number of taxa, or the number of parameters in the model grows. Nonetheless, such calculations are instructive to perform, whether to get a feel for the problem, or for developing conjectures.

EXERCISE 7.5. Consider a 2-state model of Jukes-Cantor form, with uniform root distribution and Markov matrices of the form

$$M_e = \begin{pmatrix} 1-a_e & a_e \\ a_e & 1-a_e \end{pmatrix}$$

on a 4-leaf tree. Using a leaf as a root, explicitly write down the map ϕ_T. You should have $2^4 = 16$ polynomials, expressing $p_{ijkl} = P(i,j,k,l)$ in terms of the five variables a_e. Then, using computational algebra software, find a basis for the ideal of phylogenetic invariants for this model and tree. These will be polynomials in the 16 variables p_{ijkl} found by elimination of the a_e.

The model in this last exercise, called the 2-state symmetric model or Neyman model, has as few parameters as possible to still be biologically plausible. This was in fact the model Cavender and Felsenstein worked with in [**CF87**]. To understand the difficulty of finding invariants computationally, a reader might repeat the exercise while either increasing the number of states in the model, increasing the number of taxa, or both.

There are other drawbacks to a purely computational approach to finding invariants. To perform elimination, one specifies a *term-order*, a linear ordering on monomials that induces a linear ordering on polynomials. This term-order affects the form of the results of most computations, including the computed generators of the ideal of invariants. Though one would like to understand how the model and tree topology are reflected in the form of the invariants, this may not be apparent from examining the output of a computation.

But what of non-computational approaches? How else can we find invariants? For any tree and model, the one obvious relationship between pattern frequencies is the trivial or *stochastic invariant*,

$$\sum_{\mathbf{i}\in[\kappa]^n} p_{\mathbf{i}} - 1.$$

This simply makes the claim that at any site some pattern must occur. Beyond this observation, finding invariants depends very much on both the model and the tree.

We illustrate with a few examples from [**CF87**], so we work with the 2-state symmetric model, denoting the states by 0 and 1, and consider a 4-leaf tree with neighbor pairs a, b and c, d.

First, note the 2 states are treated symmetrically, since we have a uniform root distribution and the Markov matrices are symmetric. This rather easily leads to the fact that if $\mathbf{i} \in \{0,1\}^4$ and $\mathbf{i}' = (1,1,1,1) - \mathbf{i}$ is its complement, then $p_{\mathbf{i}} - p_{\mathbf{i}'} = 0$. This gives us 8 independent linear invariants, called symmetry invariants.

To find another invariant, note that for this model, we can also develop a distance formula analogous to the Jukes-Cantor one. As the reader can show it is

$$d(x,y) = -\frac{1}{2}\ln(1-2a_{xy}),$$

where a_{xy} denotes the expected frequency of differing sites in comparing the sequences for taxa x and y. Assuming we order the four taxa as a, b, c, d, then for

instance a_{ac} can be computed by

$$a_{ac} = \sum_{i,j} p_{1i0j} + p_{0i1j}.$$

Now the 4-point condition (3.1) tells us

$$d(a,c) + d(b,d) = d(a,d) + d_(b,c).$$

So multiplying this by -2, substituting into it the formula above for the distance, and exponentiating, we get the invariant

$$\begin{aligned}(7.1)\quad &\left(1 - 2\left(\sum_{i,j} p_{1i0j} + p_{0i1j}\right)\right)\left(1 - 2\left(\sum_{i,j} p_{i1j0} + p_{i0j1}\right)\right) \\ &- \left(1 - 2\left(\sum_{i,j} p_{1ij0} + p_{0ij1}\right)\right)\left(1 - 2\left(\sum_{i,j} p_{i01j} + p_{i10j}\right)\right).\end{aligned}$$

Though this can be expressed more concisely by taking advantage of the stochastic invariant, we have established that there is a quadratic invariant that is tied to the topological structure of the tree. This polynomial vanishes only for the 4 leaf tree where a and b are neighbors, and does not vanish for generic joint distributions arising from the other two 4-leaf topologies. Invariants such as this one are said to be topologically *informative*.

In some ways the construction of this invariant, depending as it did on the particular model's distance formula and 4-point condition, was misleading in that it is simply a different presentation of a distance idea. Invariants for more complicated models, or even the other invariant for this model presented in [**CF87**], are not so tied to distance ideas. On the other hand, this invariant does express in a direct way the topology of the tree. It is highly desirable that invariants be associated to particular features, such as edges or nodes, within a tree.

After phylogenetic invariants were introduced in 1987, much work focused on linear invariants for different models. One reason for the emphasis on linear ones was the understanding that these would vanish not only on joint distributions arising from the basic model, but also on extensions of the model in which rate variation across sites was allowed. It was established for some models that linear invariants alone provided a statistically consistent method of inference. Unfortunately, simulation studies showed that using linear invariants for tree inference typically required very long data sequences to perform well in practice, much longer than other methods.

Finding invariants, especially higher degree ones, also remained difficult. Notable successes were achieved only for the group-based models. For them, Fourier analysis on the (abelian) group allowed several different collaborations [**ES93, SSEW93**] to construct invariants. The relevant Fourier transform ideas had already been introduced in the form of the Hadamard conjugation mentioned earlier. This thread was further developed in [**SS05**], where it was recognized that the change of variables associated with the Fourier transform showed the variety was toric, and thus the ideal could be fully understood.

Remark 7.6. In finding phylogenetic invariants, we'd prefer to determine the full ideal I_T. However, a weaker goal is to merely determine a set of polynomials

whose zero set is V_T. That is, we might be able to find a *set-theoretic* definition of the variety without determining a *scheme-theoretic* definition. Set-theoretic defining polynomials generate an ideal whose radical is the full ideal, but determining that radical may be difficult.

The GM Model. In the setting of the GM model, when $\kappa = 2$ *all* invariants can be understood as arising from topological features of a tree T, and for larger κ that is at least conjecturally true. We will outline some of the results from [**AR05a**] to elaborate on these claims. Note that many of the other models we have mentioned are submodels of the GM model, and so invariants for GM are also invariants for them.

First suppose the number of states for our model is $\kappa = 2$, with states denoted by 0 and 1. We give a small example, for a tree with only a few taxa, in order to clarify our notation. Consider the 5-taxon tree of Figure 7, and let $P = (p_{i_1 \cdots i_5})$ denote the joint distribution of bases at the leaves, under the GM model.

Focus on one of the internal branches of T, labeled by e in the figure. Deleting e partitions the taxa as $\{a_1, a_2\}$ and $\{a_3, a_4, a_5\}$. This partition is called the *split* induced by e. (An important combinatorial result, the *Splits Equivalence Theorem*, states that a tree is uniquely determined by its set of splits. See [**SS03**] for a proof.)

Imagine now a statistical model based on the split induced by e: Group the taxa a_1a_2, and the taxa $a_3a_4a_5$, so each is on a leaf attached to a common ancestral node. Then, the numbers of states at the leaves is 4 and 8 respectively, and we can use binary notation to denote states at the leaves. For example, the four states at leaf a_1a_2 are $00, 01, 10, 11$. Forming the joint distribution for this 'coarser' model, we get a 4×8 matrix $\mathrm{Flat}_e(P)$ given by

$$\mathrm{Flat}_e(P) = \begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} & p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} & p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} & p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} & p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}.$$

Here, for example, the $(01, 000)$-entry of $\mathrm{Flat}_e(P)$ is the probability of observing state 01 at leaf a_1a_2, and state 000 at leaf $a_3a_4a_5$. Of course, this entry is precisely p_{01000}.

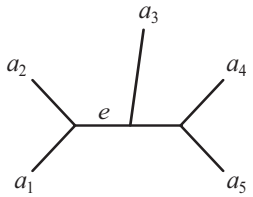

FIGURE 7. A 5-taxon tree.

The other internal edge of T similarly induces the split $\{\{a_1, a_2, a_3\}, \{a_4, a_5\}\}$ and a 'coarser' model with joint distribution given by the 8×4 matrix

$$\mathrm{Flat}_{e_2}(P) = \begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} \\ p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} \\ p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} \\ p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} \\ p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}.$$

Here the rows are indexed by the states at $a_1a_2a_3$ and columns by the states at a_4a_5.

Each of the two matrices above are simply rearrangements of the entries of the 5-dimensional tensor P into 2-dimensional arrays. Each such *flattening* is associated with a split, or internal edge, of T.

From these examples, it should be clear that for an n-leaf tree, where P is n-dimensional, we can similarly define the matrices $\mathrm{Flat}_e(P)$, where e is any edge of the tree.

THEOREM 7.7. [**AR05a**] *For the GM model with $\kappa = 2$ states on an n-leaf tree T, the phylogenetic ideal I_T is generated by all 3×3 minors of $\mathrm{Flat}_e(P)$ for all edges e of T.*

In the specific case of the 5-taxon tree of Figure 7, the theorem says all 3×3 minors of the two matrices above generate I_T. We need not bother with flattenings along pendant edges, since they have no 3×3 minors.

Notice especially that Theorem 7.7 is a scheme-theoretic statement; it says that *all* phylogenetic invariants are generated by the minors. Moreover, it relates topological features of T (edges e) to invariants (minors).

It is worthwhile to outline some ideas that arise in the proof Theorem 7.7, to gain more insight into how this result might be generalized to $\kappa > 2$, and into the special circumstances for $\kappa = 2$ that allow us to obtain a scheme-theoretic result.

We begin by examining the 'coarser' graphical models that gave rise to the flattenings of a joint distribution P. If the root is placed at either end of the edge e, then the coarser model may be depicted graphically as on the right in Figure 8. If the tree on the right is denoted by T_e, then from numerical parameters on T, it is possible to derive numerical parameters on T_e. These would be a root distribution vector $\boldsymbol{\pi}_r$ and two Markov matrices, with M_1 of size $\kappa \times \kappa^2$, and M_2 of size $\kappa \times \kappa^3$. Note that this model on T_e is *not* a phylogenetic one, since the number of states at the leaves are differing powers of κ, though there are still κ states at the root.

Specifically, if by Proposition 4.2 we assume the root r of T is located at the left end of e, then the root distribution vector $\boldsymbol{\pi}$ on T_e can be taken to be that of T. Then if M_{a_1} and M_{a_2} are the Markov matrices on the edges of T leading to a_1 and a_2 respectively, we define M_1 by $M_1(i, (j, k)) = M_{a_1}(i, j)M_{a_2}(i, k)$. The matrix M_2 is constructed similarly, but involves entries from the Markov matrices on the three other edges of T.

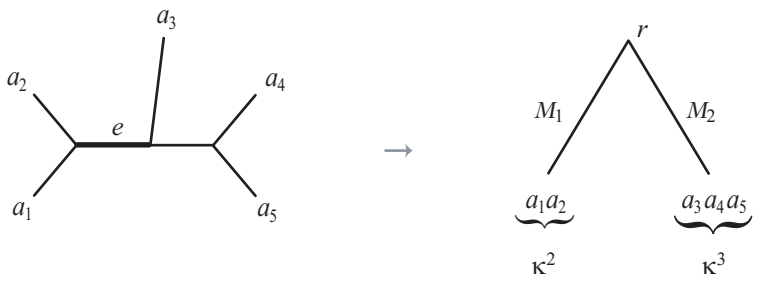

FIGURE 8. A graphical model giving rise to edge invariants.

The relationship between the joint distribution $\text{Flat}_e(P)$ for the coarser model on T_e and its numerical parameters is now expressed as:

$$\text{Flat}_e(P) = M_1^T \,\text{diag}\,(\boldsymbol{\pi}) M_2. \tag{7.2}$$

EXERCISE 7.8. Verify Equation (7.2).

Equation (7.2) immediately reveals why phylogenetic invariants arising from edge flattenings must exist for any tree T and any number of states κ. Since the diagonal matrix $\text{diag}\,(\boldsymbol{\pi})$ is of size $\kappa \times \kappa$, $\text{Flat}_e(P)$ must have rank $\leq \kappa$. The edge invariants, $(\kappa+1) \times (\kappa+1)$-minors from flattenings along edges, are phylogenetic invariants indicating that $\text{Flat}_e(P)$ satisfies the given rank condition.

REMARK 7.9. The well-known fact that the vanishing of all $(k+1) \times (k+1)$ minors of a matrix implies its rank is at most k ensures these minors are in the ideal defining the variety of rank $\leq k$ matrices. In fact, these minors generate that ideal.

DEFINITION 7.10. Suppose T is an n-taxon tree with κ states at each node, and P a joint distribution of states at the leaves of T arising from the GM model on T, or any submodel of the GM model. Let $\text{Flat}_e(P)$ denote the flattening of P induced by an edge e of T. Then the collection of $(\kappa+1) \times (\kappa+1)$-minors of $\text{Flat}_e(P)$ is the set of *edge invariants for* e. The set of *edge invariants of* T is the union of the sets of edge invariants for all edges of T.

We therefore have shown

PROPOSITION 7.11. *For any κ, the κ-state GM model on T, or any submodel, the phylogenetic ideal contains all edge invariants.*

Theorem 7.7 thus claims that edge invariants are essentially the only invariants for GM when $\kappa = 2$. This was conjectured in [**PS04b**].

Though the construction of edge invariants is natural from the viewpoint of statistical models, the proof of Theorem 7.7 involves different sorts of mathematical ideas: a special fact about a certain Segre variety when $\kappa = 2$, group actions of $GL(2)$ and $GL(4)$ on varieties, and some representation theory.

To hint at this material, we explain a connection between the κ-state GM model on a 3-leaf tree and a classical object in algebraic geometry. More details can be found in [**GSS05, AR05a**].

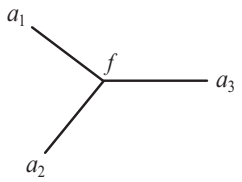

FIGURE 9. The 3-taxon tree.

In Section 4, when stochastic models of the base substitution process were introduced, we assumed that the root distribution vectors and rows of the Markov matrices sum to 1. Indeed, the probabilistic interpretation of our model required that. However, from a viewpoint of algebraic geometry, these conditions simply say the row vectors are chosen from a certain affine subset of a projective space. If we use projective coordinates, so vectors are determined only up to scalar multiples, we view each row of the Markov matrices as an element of $\mathbb{P}^{\kappa-1}$.

At the same time, we should view V_T projectively. The stochastic invariant, which states that the entries of $P \in V_T$ add to 1, tells us V_T actually lies in an affine subset of $\mathbb{P}^{\kappa^n-1}$. A projective viewpoint means we drop the stochastic invariant, and look for generators of a homogeneous ideal of phylogenetic invariants.

Consider then the 3-taxon tree T_3 of Figure 9 in the projective setting for κ states. Fix the root at the internal node f of T_3 and suppose, momentarily, that the state at the root is fixed as l. Then for each edge leading away from f, towards taxon a_i, we have a point $\boldsymbol{v}_{la_i} \in \mathbb{P}^{\kappa-1}$ that represents the lth row of a Markov matrix. The entries of $\boldsymbol{v}_{la_i} = (v_{l1}, \cdots, v_{lj}, \cdots, v_{l\kappa})$ denote, up to a scaling factor, the probability that state l at f becomes state j at a_i.

Thus, if we form

$$P^l = \boldsymbol{v}_{la_1} \otimes \boldsymbol{v}_{la_2} \otimes \boldsymbol{v}_{la_3} \in \mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1},$$

then P^l is a point in the Segre product of three projective spaces whose entries (up to scaling) are the expected frequency of observing pattern ijk conditioned on the state at the internal node f being l.

Summing over all possible states at f, we obtain the joint distribution P is

$$P = P^1 + P^2 + \cdots + P^\kappa.$$

(While not explicitly appearing, the root distribution has been accounted for in the arbitrary scaling factors that appear in each P^l when we choose particular projective coordinates to express them.) Now just as sums of two points on a projective variety gives points on secant lines to the variety, the (closure of the) union of which is the *secant variety*, we can consider higher secant varieties as well. Since we are summing κ points on the variety $\mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1}$, we obtain

$$P \in V_{T_3} = \mathrm{Sec}^\kappa(\mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1}),$$

the κ-secant variety of the Segre product of three $\mathbb{P}^{\kappa-1}$.

More concretely, the joint distribution P has been decomposed as the sum of κ *rank 1* tensors, one for each possible state at the internal node f. This is precisely the definition that P has *tensor rank* at most κ, and we have established that the phylogenetic variety V_{T_3} is the (closure of) the set of $\kappa \times \kappa \times \kappa$ tensors of rank at most κ.

The concept of tensor rank, as the minimal number of rank 1 summands to produce a tensor, parallels one of the many possible definitions of matrix rank. For those who have not run across tensor rank before, we point out it is considerably more subtle than its matrix analogue. For instance, there is still no straightforward way to determine the rank of an arbitrarily chosen tensor, even in the 3-dimensional case. Neither analogues of matrix minors, nor an algorithmic method analogous to Gaussian elimination are known. However, because of the widespread appearance of the concept in applications, there are approaches to finding decompositions as sums of rank 1 tensors, though not necessarily minimal ones.

When $\kappa = 2$, however, things are simple. Indeed, the GM model on a 3-taxon tree has only 7 parameters, and since the stochastic invariant cuts out a 7-dimensional subspace of $\mathbb{C}^{2^3}$, one might conjecture there are no other invariants. In fact, this is the case, and $\mathrm{Sec}^2(\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1) = \mathbb{P}^7$. In other words, every $2 \times 2 \times 2$ tensor is in the closure of the rank 2 ones. (Note this does not mean every such tensor has rank 2.) This special fact plays an important role in [**AR05a**].

For $\kappa = 3$, the ideal defining $\mathrm{Sec}^\kappa(\mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1})$ was found in [**GSS05**], using results from [**Str83**]. For $\kappa = 4$, polynomials are known that generate the ideal only up to saturation with respect to another explicitly given variety, and then taking a radical [**AR03**]. In this case there are 1728 independent quintics, which are known to be all such quintics. See [**Hag00**] for computation of this dimension, or [**LM04**] for a broader set of computations of dimensions of spaces of polynomials vanishing on various secant varieties of Segre varieties.

Note also that for the 3-taxon tree, the construction of edge invariants yields nothing, since there are no internal edges. This shows that any hope that edge invariants might generate the ideal for $\kappa > 2$ fails even for the 3-taxon tree. At the same time, however, the existence of 3-taxon invariants suggests a path to understanding I_T, for arbitrary trees T, through *vertex flattenings*.

More specifically, for an arbitrary tree, focus on a node v and *flatten* to a 'coarser' graphical model, as shown in Figure 10. Correspondingly, for the n-dimensional joint distribution $P \in V_T$, flatten

$$P \mapsto \mathrm{Flat}_v(P),$$

where $\mathrm{Flat}_v(P)$ is a $\kappa^{n_1} \times \kappa^{n_2} \times \kappa^{n_3}$ tensor, $n_1 + n_2 + n_3 = n$, obtained by grouping taxa as in the tree. Again, the coarsening and flattening operations focus attention on a local feature of T, in this case, the tri-partitioning of the n taxa that the node v implies.

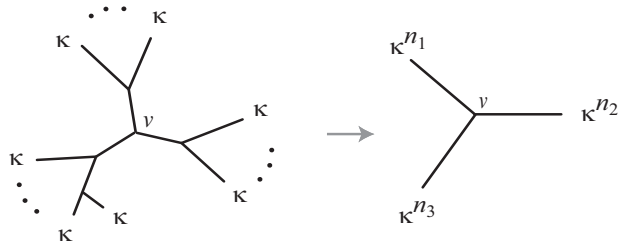

FIGURE 10. A vertex flattening of a model.

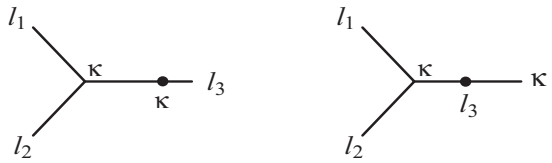

FIGURE 11. 'Extending the edge' gives rise to maps between varieties.

Now the variety associated to the coarsened model is $\mathrm{Sec}^{\kappa}(\mathbb{P}^{\kappa^{n_1}-1} \times \mathbb{P}^{\kappa^{n_2}-1} \times \mathbb{P}^{\kappa^{n_3}-1})$, the variety of rank κ tensors of size $\kappa^{n_1} \times \kappa^{n_2} \times \kappa^{n_3}$.

Our next step is to develop the relationship between the varieties $\mathrm{Sec}^{\kappa}(\mathbb{P}^{\kappa^{n_1}-1} \times \mathbb{P}^{\kappa^{n_2}-1} \times \mathbb{P}^{\kappa^{n_3}-1})$ and $\mathrm{Sec}^{\kappa}(\mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1})$. For this purpose, we introduce the notation

$$V(\kappa; l_1, l_2, l_3) = \mathrm{Sec}^{\kappa}(\mathbb{P}^{l_1-1} \times \mathbb{P}^{l_2-1} \times \mathbb{P}^{l_3-1})$$

for the variety for the model on the 3-leaf tree with κ states at the internal node, and l_1, l_2, l_3 states at the leaves.

For any joint distribution $P \in V(\kappa; \kappa, \kappa, \kappa)$ observe that there is an 'action' by $\kappa \times l_3$ complex matrices M in the third index of P. In modeling language, we think of this action as 'extending the edge' leading to the third leaf by tacking on an additional state-change process represented by the matrix M, as shown on the left in Figure 11. This gives us a point $P *_3 M \in V(\kappa; \kappa, \kappa, l_3)$. In terms of parameters, if $P = \phi_T((\boldsymbol{\pi}, \{M_1, M_2, M_3\}))$ where M_3 is the matrix on the edge leading to the third leaf, then $P *_3 M = \phi_T(\boldsymbol{\pi}, \{M_1, M_2, M_3 M\})$, though the action extends to the points on the variety that are not in the image of ϕ_T as well.

We may similarly define an 'action' of $l_3 \times \kappa$ matrices N on $V(\kappa; \kappa, \kappa, l_3)$, as depicted on the right in Figure 11. Then, for every choice of $\kappa \times l_3$ matrix M and $l_3 \times \kappa$ matrix N, we have maps

$$V(\kappa; \kappa, \kappa, \kappa) \underset{*_3 N}{\overset{*_3 M}{\rightleftarrows}} V(\kappa; \kappa, \kappa, l_3).$$

These maps give rise to corresponding maps between the ideals defining the varieties, whose compositions are related to $GL(\kappa)$-and $GL(l_3)$-actions. With this setup, a careful use of basic representation theory gives the following important result.

THEOREM 7.12. [**AR05a**] *If $l_i \geq \kappa$, and S is any set of polynomials defining $V(\kappa; \kappa, \kappa, \kappa)$ set-theoretically (resp., scheme-theoretically), then from S an explicit set of polynomials defining $V(\kappa; l_1, l_2, l_3)$ set-theoretically (resp., scheme-theoretically) can be constructed.*

Because Theorem 7.12 relates phylogenetic invariants on T_3 to sets of polynomials defining the varieties $V(\kappa; \kappa^{n_1}, \kappa^{n_2}, \kappa^{n_3})$ that appear in vertex flattenings, it is one of the needed ingredients to determine a set-theoretic description of the phylogenetic variety V_T for the general Markov model on any n-taxon tree T. We state the resulting theorem somewhat informally.

THEOREM 7.13. [**AR05a**] *For the 3-taxon tree T_3, let S be a set of polynomials defining $V(\kappa;\kappa,\kappa,\kappa)$ set-theoretically. Then, using vertex flattenings and the construction of Theorem 7.12, for an arbitrary binary tree T a set of polynomials set-theoretically defining V_T for the general Markov model can be explicitly given.*

An important consequence of Theorem 7.13 is that phylogenetic invariants for the general Markov model are intimately related to the nodes and edges of T. The local structure of a tree determines a collection of phylogenetic invariants defining the variety V_T. Note that (using different techniques) this sort of result for group-based models had already been established in [**SS05**]. Possible ways this might be useful will be discussed in the next section.

8. Potential uses of invariants

So far, invariants have not planned a large role in practical inference by biologists. However, now that we are beginning to understand them better, that may well change. In this section we outline some of the ideas now under development.

Tree-building heuristics. A key property of the invariants we understand is that specific polynomials can be tied to local structure of a tree (edges or nodes). They might therefore be used to develop tests for *only* such local structures, without consideration of the entire tree.

To elaborate, one inherent feature of maximum likelihood is that it not only chooses the 'best' tree, but also 'best' values for all parameters. This is precisely why ML inference can be such a large computational problem; it looks at everything at once. However, for building a tree from data (and for some biological questions) we might first look for strongly-supported splits of our taxa (in biological terminology, for monophyletic clades). If specific invariants can be tied to edges (splits) and nodes (tripartitions), perhaps we can address the support for each feature individually.

One step toward using this viewpoint to infer trees has been taken in [**Eri05**]. There an algorithm is given for building trees that is reminiscent of Neighbor Joining in its 'outside-first' iterative approach. The scheme for joining taxa, though, is based on edge invariants. Rather than evaluate polynomials, however, the singular value decomposition of matrices is used to determine approximate matrix rank. Preliminary results on the algorithm's performance were reported as positive, though not as strong as more standard methods. Nonetheless, the comparisons were probably biased *against* the new method, since data was simulated according to much simpler model than GM, that, among other things, assumes the same distribution of bases in all sequences. We believe that there is much room for further development along these lines.

Even if we prefer to stick with a full maximum likelihood framework for inference, we must acknowledge that implementations in software require heuristic searches if more than a handful of taxa are involved. For the numerical parameters, optimization is a well-studied problem and we might assume this part of the search can be done adequately. For the tree parameter, though, how should we vary the tree in order to increase likelihoods? Perhaps invariants can be used to identify more weakly supported edges or nodes in the tree which should be removed in reconfiguring. If they are effective at suggesting how we might move toward optima in tree-space, they may help speed up searches.

Exact solutions of ML problems. Since for large problems, ML inference must be done heuristically, it would be desirable to understand better under what circumstances there might be one, or more, global optimum, and whether we have many local optima. Invariants have placed a role in studying these questions, by allowing ML estimation to be phrased as a constrained optimization problem, with invariants providing the constraints. See [**CHHP00, CHP01, CKS03**] for some examples of this sort of work. [**HKS05**] provides a more general setting for computational algebra approaches to ML, as well as phylogenetic examples.

The "Small Trees website' [**CGS05**] is a good resource allowing an easy interface from trees and models to computational algebra package formulations. Finally [**CL05**] suggests how solving ML problems well on small trees can, with a generalized Neighbor Joining approach, lead to construction of large trees.

Identifiability of tree topologies for models. An important issue in dealing with any statistical model is *identifiability*: Given a joint distribution arising from the model, is it possible to recover the parameters leading to the distribution? Clearly, identifiability of any parameters of interest is a necessary condition to our estimating them well. Indeed, proofs of the statistical consistency of ML begin with proofs of identifiability.

Identifiability has been established for many phylogenetic models routinely used (for instance, GTR+I+Γ; GM, and hence any submodel of GM). Provided a distance can be defined for the model, the 4-point condition can be used to identify topologies. In fact, since distances require comparing only two sequences at a time (i.e., are based on 2-dimensional marginalizations of the joint distribution P), identifying the tree does not even require the full joint distribution. On the other hand, [**Baa98**] established that the tree could *not* be identified by 2-sequence comparisons for the model GM+I. In general, identifiability for mixture models of this sort has been poorly understood. Even for GTR+(rate distribution), proofs of identifiability of the tree require that the rate distribution be known. See [**SSH94**] and [**BGP05**].

Recently some specific invariants that have not been discussed here have been used to obtain some general results on identifiability of tree topologies. In [**AR05b**] it is shown that for a mixture model where the number of classes is less than the number of states (e.g., at most 3 classes for a 4-state model for DNA), tree topology is identifiable for generic choices of parameters. This result makes no assumptions of any commonality to the substitution process among the different classes; they need not be based on any common rate matrix. The result also applies to a *covarion* model [**TS98**] in which sites in a sequence may only be free to vary in some (unknown) parts of the tree, switching between variable and invariable and back as evolution proceeds over the tree

To prove such identifiability results, an algebraic mutation model is introduced which allows more states at internal nodes of the tree than at the leaves. In this very general setting, it is possible to show that parameters are generically identifiable, and then to argue that each of the models listed above is a specialization of this model, and that generic identifiability is maintained after specialization. The key steps in the proof involve finding invariants that express appropriate rank conditions similar to those that arose in the edge flattenings and vertex flattenings of Theorems 7.7 and 7.13.

9. Further Reading and Software Packages

For mathematicians interested in learning more about phylogenetics, we suggest a few books. Felsenstein's book [**Fel04**], written for a diverse audience, is the most complete survey of the field, with many references. The text [**SS03**] by Semple and Steel is more mathematical, emphasizing the combinatorial aspects. Pachter and Sturmfels, in [**PS05**], provide an excellent introduction to the algebraic viewpoint. The volume [**Gas05**] provides a collection of articles on many more aspects of phylogenetics than we have been able to even touch on here.

For undergraduate teaching, the only presentation we know of any material on the mathematics behind phylogenetics appears in [**AR04**].

There are many software packages used by biologists for phylogenetic inference. PHYLIP, which is freely available over the web from the Felsenstein lab is a good collection of programs to begin exploring. Most published biological papers will indicate what software was used for inference, so finding pointers to other packages is relatively easy.

References

[AR03] Elizabeth S. Allman and John A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186:113–144, 2003.

[AR04] Elizabeth S. Allman and John A. Rhodes. *Mathematical Models in Biology: an Introduction*. Cambridge University Press, 2004.

[AR05a] Elizabeth S. Allman and John A. Rhodes. Phylogenetic ideals and varieties for the general Markov model. 2005. preprint, `arXiv:math.AG/0410604`.

[AR05b] Elizabeth S. Allman and John A. Rhodes. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. 2005. preprint, `q-bio.PE/0511009`.

[AR06] Elizabeth S. Allman and John A. Rhodes. Phylogenetic invariants for stationary base composition. *J. Symbolic Comp.*, 41(2):138–150, 2006.

[Baa98] Ellen Baake. What can and what cannot be inferred from pairwise sequence comparisons? *Math. Biosci.*, 154(1):1–21, 1998.

[BGP05] David Bryant, Nicolas Galtier, and Marie-Anne Poursat. Likelihood caclulation in molecular phylogenetics. In Olivier Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 33–62. Oxford University Press, 2005.

[Bun71] Peter Buneman. The recovery of trees from measures of dissimilarity. In *Mathematics in the Archeological and Historical Sciences*, pages 387–395, Edinburgh, 1971. Edinburgh University Press.

[CF87] James A. Cavender and Joseph Felsenstein. Invariants of phylogenies in a simple case with discrete states. *J. of Class.*, 4:57–71, 1987.

[CGS05] Marta Casanellas, Luis David Garcia, and Seth Sullivant. Catalog of small trees. In Lior Pachter and Bernd Sturmfels, editors, *Algebraic Statistics for Computational Biology*, pages 291–304. Cambridge University Press, 2005. `http://www.math.tamu.edu/~lgp/small-trees/`.

[CHHP00] B. Chor, M. D. Hendy, B. R. Holland, and D. Penny. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Mol. Bio. and Evol.*, 17:1529–1541, 2000.

[CHP01] Benny Chor, Michael Hendy, and David Penny. Analytic solutions for three-taxon ML_{MC} trees with variable rates across sites. In *Algorithms in bioinformatics (Århus, 2001)*, volume 2149 of *Lecture Notes in Comput. Sci.*, pages 204–213. Springer, Berlin, 2001.

[CKS03] B. Chor, A. Khetan, and S. Snir. Maximum likelihood on four taxa phylogenetic trees: Analytic solutions. *RECOMB'03*, 2003.

[CL05] Mark Contois and Dan Levy. Small trees and generalized neighbor-joining. In Lior Pachter and Bernd Sturmfels, editors, *Algebraic Statistics for Computational Biology*, pages 335–346. Cambridge University Press, 2005.

[CLO97] David Cox, John Little, and Donal O'Shea. *Ideals, varieties, and algorithms*. Springer-Verlag, New York, second edition, 1997.

[Eri05] Nicholas Eriksson. Tree construction using singular value decomposition. In Lior Pachter and Bernd Sturmfels, editors, *Algebraic Statistics for Computational Biology*, pages 347–358. Cambridge University Press, 2005.

[ES93] Steven N. Evans and T. P. Speed. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.*, 21(1):355–377, 1993.

[Fel78] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27:401–410, 1978.

[Fel04] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA, 2004.

[FG82] L. R. Foulds and R. L. Graham. The Steiner problem in phylogeny is NP-complete. *Adv. in Appl. Math.*, 3(1):43–49, 1982.

[Gas05] Olivier Gascuel, editor. *Mathematics of Evolution and Phylogeny*. Oxford University Press, Oxford, 2005.

[GPS01] G.-M. Greuel, G. Pfister, and H. Schönemann. SINGULAR 2.0. A Computer Algebra System for Polynomial Computations, Centre for Computer Algebra, University of Kaiserslautern, 2001. `http://www.singular.uni-kl.de`.

[GS02] Daniel R. Grayson and Michael E. Stillman. Macaulay2, a software system for research in algebraic geometry. Available at `http://www.math.uiuc.edu/Macaulay2/`, 2002.

[GSS05] Luis David Garcia, Michael Stillman, and Bernd Sturmfels. Algebraic geometry of Bayesian networks. *J. Symbolic Comp.*, 39:331–355, 2005. `arXiv:math.AG/0301255`.

[Hag00] Thomas R. Hagedorn. A combinatorial approach to determining phylogenetic invariants for the general model, 2000. Technical report, Centre de recherches mathmatiques.

[Hen89] Michael D. Hendy. The relationship between simple evolutionary tree models and observable sequence data. *Systematic Zoology*, 38:310–321, 1989.

[HGH88] K. Hayasaka, T. Gojobori, and S. Horai. Molecular phylogeny and evolution of primate mitochondrial DNA. *Mol. Biol. Evol.*, 5:626–644, 1988.

[HKS05] Serkan Hoşten, Amit Khetan, and Bernd Sturmfels. Solving the Likelihood Equations. *Found. Comput. Math.*, 2005. to appear, `arXiv:math.ST/0408270`.

[HP89] Michael D. Hendy and David Penny. A framework for the quantitative study of evolutionary trees. *Systematic Zoology*, 38:297–309, 1989.

[Lak87] J.A. Lake. A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Bio. Evol.*, 4(2):167–191, 1987.

[LM04] J. M. Landsberg and L. Manivel. On the ideals of secant varieties of Segre varieties. *Found. Comput. Math.*, 4(4):397–422, 2004.

[PS04a] L. Pachter and D. Speyer. Reconstructing trees from subtree weights. *Appl. Math. Lett.*, 17(6):615–621, 2004.

[PS04b] Lior Pachter and Bernd Sturmfels. Tropical geometry of statistical models. *Proc. Natl. Acad. Sci. USA*, 101(46):16132–16137 (electronic), 2004.

[PS05] Lior Pachter and Bernd Sturmfels, editors. *Algebraic Statistics for Computational Biology*. Cambridge University Press, Cambridge, 2005.

[SK88] J.A. Studier and K.J. Keppler. A note on the neighbor-joining algorithm of saitou and nei. *Mol. Biol. Evol.*, 5(5):729–731, 1988.

[SN87] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.

[SS03] Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[SS05] Bernd Sturmfels and Seth Sullivant. Toric ideals of phylogenetic invariants. *J. Comput. Biol.*, 12(2):204–228, 2005. `arXiv:q-bio.PE/0402015`.

[SSE93] L. A. Székely, M. A. Steel, and P. L. Erdős. Fourier calculus on evolutionary trees. *Adv. in Appl. Math.*, 14(2):200–210, 1993.

[SSEW93] Mike Steel, Laszlo Székely, Peter L. Erdös, and Peter Waddell. A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. *N.Z. J. Botany*, 31(31):289–296, 1993.

[SSH94] M.A. Steel, L. Székely, and M.D. Hendy. Reconstructing trees from sequences whose sites evolve at variable rates. *J. Comput. Biol.*, 1(2):153–163, 1994.

[Str83] V. Strassen. Rank and optimal computation of generic tensors. *Linear Algebra Appl.*, 52/53:645–685, 1983.

[TS98] Chris Tuffley and Mike Steel. Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.*, 147(1):63–91, 1998.

Department of Mathematics and Statistics, University of Alaska Fairbanks, PO Box 756660, Fairbanks, Alaska 99775
E-mail address: e.allman@uaf.edu

Department of Mathematics and Statistics, University of Alaska Fairbanks, PO Box 756660, Fairbanks, Alaska 99775
E-mail address: j.rhodes@uaf.edu