Statistica Sinica **17**(2007), 1299-1316

# MOLECULAR PHYLOGENETICS FROM AN ALGEBRAIC VIEWPOINT

Elizabeth S. Allman and John A. Rhodes

*University of Alaska, Fairbanks*

*Abstract:* The probabilistic models used in the inference of phylogenetic trees from molecular data are particularly rich in algebraic structure. This was first noticed 20 years ago when certain polynomials called phylogenetic invariants were introduced by Cavender and Felsenstein, and by Lake. Recently, however, there have been considerable advances in our algebraic understanding of these models. We survey some of this work, indicating both how algebra has been exploited for theoretical understanding and the preliminary steps that have been taken toward its use in practical inference.

*Key words and phrases:* Markov models on trees, molecular evolution, phylogenetic invariants, phylogenetics.

## 1. The Problem of Phylogenetic Inference

Phylogenetics is concerned with the inference of evolutionary relationships among a collection of organisms, or *taxa*. Most often the relationships that are sought will be described by a *phylogenetic tree.* In such a tree, each given taxon will appear at a leaf, while (unlabeled) internal nodes represent inferred ancestors.

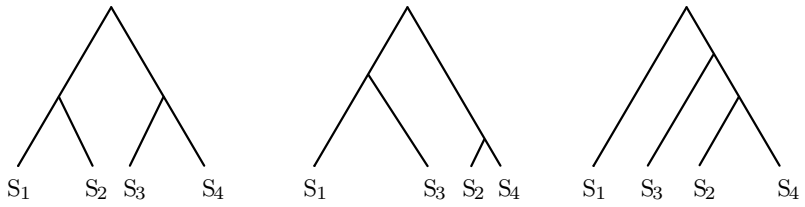


Figure 1.1. Three of the 15 possible rooted phylogenetic trees relating 4 taxa.

In Figure 1.1, for example, we see three rooted phylogenetic trees that might relate the four taxa $S_1$, $S_2$, $S_3$ and $S_4$. The roots of these trees indicate the location of the most recent common ancestor of the given taxa. However, even an unrooted tree can carry much information about evolutionary relationships,

and inference of the root location may not even be possible. Note that if roots are ignored then the two trees on the right of Figure 1.1 are topologically the same, while the one on the left is different.

Although the data underlying phylogenetic inference might be of many types, in this article we limit ourselves to discussing approaches suitable for biological sequence data (e.g., DNA, proteins). In this setting, not only can reasonable probabilistic models be devised to describe evolutionary changes, but the models also naturally display a rich algebraic structure.

Combining sequence data with an appropriate model of evolution enables phylogenetic inference to be performed in standard statistical frameworks, such as the maximum likelihood or Bayesian paradigms. Indeed, software implementations of these methods (e.g., PAUP* (Swofford (2002)), Phylip (Felsenstein (2004b)), MrBayes (Ronquist and Huelsenbeck (2003))) are widely used. However, largely because of the rapid growth in the number of possible trees that might relate a collection of taxa (for $n$ taxa, there are $(2n-3)!!$ possible rooted binary trees), the necessary computations can easily exceed what is possible. Often compromises are made, through the use of simple models and heuristic searches. Thus there remains a need for new perspectives and insights, both to better understand the nature of the problem, and to develop improved practical approaches. Algebraic statistics provides one such perspective.

While the idea of applying algebra to phylogenetic inference first emerged 20 years ago (Cavender and Felsenstein (1987) and Lake (1987)), recently there has been a resurgence of activity and progress. In subsequent sections, we introduce phylogenetic models, emphasizing their algebraic aspects. We discuss how this perspective has led to deeper theoretical understanding of the inference problem, and describe steps taken toward incorporating it into practical inference. We sketch results for those evolutionary models that are best understood, emphasizing the way tree topology is reflected in algebraic structures. We particularly hope to interest more researchers in investigating how algebraic understanding might be further exploited in data analysis.

## 2. Models of Molecular Evolution

Sequence data for a phylogenetic inference problem is usually a collection of *aligned* sequences, one for each taxon to be related. In an alignment of sequences, sites (positions) are matched so that each is presumed to correspond to an ancestral sequence site. We assume that all unalignable sites are omitted from our data so that, in the case of DNA for example, it might appear as in Table 2.1. A site will be referred to as a *character*, and the particular sequence building-block (e.g., $A$, $G$, $C$ or $T$ for DNA) in that site as a character *state*.

To introduce a simple probabilistic model of sequence evolution along a tree, consider a rooted tree, such as that in Figure 2.2, and assume that characters have the four states appropriate for DNA ($1 = A$, $2 = G$, $3 = C$, $4 = T$). We model evolution of a single character, treating the different characters in the data as independent trials of the same process (i.i.d. assumption).

Table 2.1. Aligned DNA sequences for four taxa.

| | |
|---|---|
| Taxon $S_1$: | `AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCATTATTA...` |
| Taxon $S_2$: | `AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCTCATTACTA...` |
| Taxon $S_3$: | `AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCATTATTA...` |
| Taxon $S_4$: | `AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCCCTACTG...` |

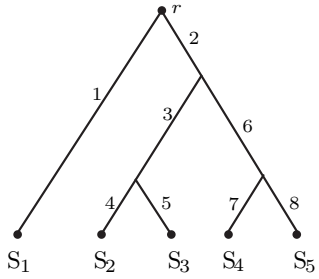


Figure 2.2. A rooted 5-leaf tree.

At the root $r$ of the tree, a *root distribution vector* $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$ describes the probabilities of each possible state occurring in the most recent common ancestor of the given taxa. For the $i$th edge of the tree, a $4 \times 4$ Markov matrix $M_i$ gives conditional probabilities of the 16 possible state changes that might occur in passing from ancestor to descendant along that edge. Thus there are $3 + 12|E|$ independent numerical parameters for this *general Markov model* of DNA evolution, where $|E| = 8$ is the number of edges in the tree. Letting $P(i_1, i_2, i_3, i_4, i_5)$ denote the probability of observing state $i_j$ at taxon $S_j$, $j = 1, \ldots, 5$, for the tree of Figure 2.2 we obtain the formula

$$\begin{aligned} &P(i_1, i_2, i_3, i_4, i_5) \\ &\quad = \sum_{s=1}^{4} \pi_s M_1(s, i_1) \left( \sum_{t=1}^{4} M_2(s, t) \left( \sum_{u=1}^{4} M_3(t, u) M_4(u, i_2) M_5(u, i_3) \right) \right. \\ &\qquad \left. \cdot \left( \sum_{v=1}^{4} M_6(t, v) M_7(v, i_4) M_8(v, i_5) \right) \right). \end{aligned} \tag{2.1}$$

The full joint distribution $P$ for an $n$-leaf tree is thus specified by an $n$-dimensional $4 \times \cdots \times 4$ table, whose entries can be calculated efficiently by polynomial formulas analogous to equation (2.1).

That formulas such as equation (2.1) are polynomial indicates that algebraic methods might be useful in phylogenetics. In fact, such polynomials are highly-structured and encode the topology of the tree; a different tree topology leads to different polynomial formulas. We refer to any model for which the joint distribution is given by polynomial formulas as *algebraic*.

Since the general Markov model is parameter-rich, more restrictive submodels can be of interest. In particular, the *Jukes-Cantor, Kimura* 2*-parameter,* and *Kimura* 3*-parameter* models all assume a uniform root distribution, and that the Markov matrices on each edge have the respective forms

$$M_{JC}=\begin{pmatrix} * & a & a & a \\ a & * & a & a \\ a & a & * & a \\ a & a & a & * \end{pmatrix},\ M_{K2}=\begin{pmatrix} * & a & b & b \\ a & * & b & b \\ b & b & * & a \\ b & b & a & * \end{pmatrix},\ M_{K3}=\begin{pmatrix} * & a & b & c \\ a & * & c & b \\ c & b & * & a \\ b & c & a & * \end{pmatrix}.$$

Here the parameters $a, b, c$ in the matrices may vary from edge to edge of the tree. The Jukes-Cantor model thus states that on each edge all possible state changes have the same probability of occurring, while the Kimura 2-parameter model allows for transitions ($A \leftrightarrow G$, $C \leftrightarrow T$) and transversions to occur with differing probabilities, as is often observed. Note that with any of these models we retain polynomial formulas for the joint distribution, such as in equation (2.1), so these models are also algebraic.

In current practical inference it is most common to assume that all Markov matrices on the tree edges arise from a common *rate matrix* $Q$. Here $Q$ is a $4 \times 4$ matrix with non-negative off-diagonal entries — the instantaneous rate of various state changes — and with rows summing to 0. Each edge $e$ of the tree is then assigned an edge length parameter $t_e$, representing actual time or some non-clock-like measure of mutation. The Markov matrix giving state change probabilities from one end of the edge $e$ to the other is then $M_e = \exp(t_e Q)$. The root distribution $\boldsymbol{\pi}$ is generally taken to be an eigenvector of $Q$ with eigenvalue 0, so that the state distribution is stationary under the substitution process on each edge, and $Q$ is assumed to be time-reversible (see Felsenstein (2004a)). Note that this parameterization is not purely algebraic, as it involves the exponential function. However, algebraic methods can be used to study these models also, as they are more restrictive submodels of the algebraic general Markov model.

The assumption of a common rate matrix across a tree is certainly dubious for some data sets, as are the stationary base distribution and time-reversible assumptions. As phylogenetic work advances and more complex data sets are collected and analyzed, it becomes desirable to relax such overly simplistic assumptions. Thus the discrete time (one matrix per edge) formulation of the algebraic models we focus on can be viewed as a potential strength, as it allows some weakening of questionable hypotheses.

Finally, it is also common to consider models in which the i.i.d. assumption is relaxed somewhat so that different sites may evolve at different rates. This may be handled by explicitly partitioning the data (if one has information on natural ways to do this) or by considering mixture models in which the partitioning is controlled by parameters. In the continuous-time formulation, one might imagine several classes of sites, each with its own scalar rate parameter $\lambda_i$, so that for class $i$ the Markov matrices are $M_e = \exp(t_e \lambda_i Q)$. The mixing parameters $s_i$ then describe the proportion of sites in each class, and the joint distribution is simply a sum, weighted by the $s_i$, of the joint distributions for each class.

For algebraic models, such as the general Markov, we can similarly construct mixture models allowing rate variation across sites. If we allow several classes of sites, then each class has its own set of general Markov parameters (with the same tree), and the joint distribution for the model is simply a weighted sum of joint distributions for each class, with weights given by the mixing parameters. For DNA evolution, such a general Markov model mixture with $N$ classes would thus have $(N-1) + N(3 + 12|E|)$ independent parameters, where $|E|$ is the number of edges in the tree.

## 3. Phylogenetic Invariants

The origin of an algebraic approach to phylogenetic inference can be traced to two independent papers (Lake (1987), Cavender and Felsenstein (1987)). (Although Lake (1987) seems to have had the greater impact in the early days, we recommend Cavender and Felsenstein (1987) for its more illuminating viewpoint and closer connection to recent work.) These papers introduced polynomials referred to as *phylogenetic invariants*, and proposed their use in inference.

Fix a phylogenetic tree $T$ with $n$ taxa labeling the leaves, and a 4-state algebraic model $\mathcal{M}$ of character evolution. We then have a polynomial parameterization

$$\phi_{T,\mathcal{M}} : S \to \mathbb{R}^{4^n}$$

giving the joint distribution in terms of the numerical parameters. Here $S \subset [0,1]^L$ denotes the numerical parameter space for the model on $T$. The image $\phi_{T,\mathcal{M}}(S)$ of this map forms a piece of a 'surface' of dimension $L$ (or less) in $4^n$-dimensional space. Because the parameterization is given by polynomials, a fundamental result from algebraic geometry ensures that there will be another set of polynomials, in the $4^n$ coordinate variables of the image space, that evaluate to zero at every point of the image. Polynomials in this second set are called *phylogenetic invariants* for $T$ and $\mathcal{M}$. They give an implicit description, as a zero-set, of the collection of all joint distributions arising from the model for this fixed tree.

More formally, the set of all phylogenetic invariants for $T$, $\mathcal{M}$ is called the *phylogenetic ideal* $I_{T,\mathcal{M}}$ of the polynomial ring $\mathbb{C}[P]$, defined by

$$I_{T,\mathcal{M}} = \{f(P) \mid f(\phi_{T,\mathcal{M}}(\mathbf{s})) = 0 \text{ for all parameters } \mathbf{s} \in S\}.$$

Although $S$ is a subset of $[0,1]^L$, since $\phi_{T,\mathcal{M}}$ is polynomial, this function extends to a polynomial map on all of $\mathbb{C}^L$. Though unnatural from a statistical viewpoint, this extension to complex parameters puts us in a convenient setting to use the tools of algebraic geometry. Importantly, $I_{T,\mathcal{M}}$ is unchanged if $S$ is replaced by $\mathbb{C}^L$ in this definition.

The *phylogenetic variety* $V_{T,\mathcal{M}}$ is the set of points on which all phylogenetic invariants vanish,

$$V_{T,\mathcal{M}} = \{P \in \mathbb{C}^{4^n} \mid f(P) = 0 \text{ for all } f \in I_{T,\mathcal{M}}\}.$$

Thus the phylogenetic variety contains all (complex) joint distributions that arise from the model on $T$, as well as some additional points in the closure of that set. While we are of course most interested in points on $V_{T,\mathcal{M}}$ that are in $\phi_{T,\mathcal{M}}(S)$, even the points on $V_{T,\mathcal{M}}$ with real coordinates form a strictly larger set.

While phylogenetic invariants give all polynomial *equalities* satisfied by joint distributions arising from a model, note that Cavender and Felsenstein (1987) already recognized that polynomial *inequalities* are also satisfied by distributions arising from stochastically-meaningful parameters. In algebraic geometry it is well-understood that if one restricts from complex parameters to real ones then inequalities are generally necessary to characterize the image of a parameterization map, and the further restriction of parameters to stochastic values can lead to additional inequalities. However, determining such inequalities explicitly is a difficult problem of real algebraic geometry, and in the phylogenetic setting much needs to be accomplished in this direction. (See, however, the recent work of Stefankovic and Vigoda (2007) on linear inequalities.) In this short survey we limit our focus to invariants.

For the general Markov model, and many others, it can be shown that phylogenetic varieties are independent of the root location in the tree; that is, even though the parametrization map $\phi_{T,\mathcal{M}}$ is defined by specifying a root, the (closure of) its image depends only on the topology of the unrooted tree (Steel, Székely and Hendy (1994) and Allman and Rhodes (2003)). (This trait is shared by the non-algebraic continuous-time reversible model so heavily used in practical inference as well.) Consequently, from now on we treat all trees as unrooted.

A naive plan for using phylogenetic invariants for inference proceeds as follows. Having chosen a particular model $\mathcal{M}$, for each possible tree $T$ that might

relate the given taxa, find a collection of phylogenetic invariants. From the sequence data, compute the observed joint distribution $\widehat{P}$. Then choose as the best tree the one for which the invariants are somehow 'closest' to zero when evaluated on $\widehat{P}$.

In following such a scheme, we emphasize that there is no estimation of any numerical model parameters. Whether one views such parameters as 'nuisances,' or merely recognizes that their inference can be time-consuming, this is one reason why invariants are attractive. Of course, with a tree in hand inference of numerical parameters can then be performed more quickly.

In principle, generators of the phylogenetic ideal can be computed from the polynomials defining the parameterization map, using computational algebra techniques built on Gröbner bases. In practice, however, except for the simplest models and small trees, the number of variables involved in such a calculation is too great for even the best current software.

To give concreteness to the idea of an invariant, we sketch two constructions. The first of these comes from Cavender and Felsenstein (1987); the second is also motivated there, and plays an important role in more recent works (Sturmfels and Sullivant (2005), Allman and Rhodes (2007a)). Both illustrate the important point that specific invariants can often be given very natural interpretations as statements about a tree and model.

Consider the 4-taxon tree of Figure 3.3 viewed as a metric tree, and let $d_{ij}$ denote the total edge-length distance along the tree between taxa $\mathrm{S}_i$ and $\mathrm{S}_j$. Then the *4-point condition* (Buneman (1971)) asserts

$$d_{12} + d_{34} < d_{13} + d_{24} = d_{14} + d_{23}, \tag{3.1}$$

for any assignment of positive edge lengths to this tree. Furthermore, the two other unrooted binary topological trees that might relate four taxa lead to distances that do not satisfy this condition. Now for the general Markov model on the tree of Figure 3.3, let $P_{ij}$ denote the matrix of expected frequencies of pairs of states at the taxa $\mathrm{S}_i$ and $\mathrm{S}_j$ (so $P_{ij}$ is a 2-dimensional marginalization of the full 4-dimensional table $P$ giving the joint distribution). Then one can show that $D_{ij} = \det P_{ij}$ behaves much like a multiplicative version of a distance on the tree, and thus the equality (3.1) yields, for a DNA model, the degree 8 polynomial equation

$$D_{13}D_{24} - D_{14}D_{23} = 0.$$

Indeed, for the general Markov model, $D_{ij}$ is (up to some minor adjustments) the exponential of the log-det distance later introduced by Steel (1994). Since the Jukes-Cantor and Kimura models are submodels of the general Markov one, the left hand side of this equation is an invariant for all of these models.

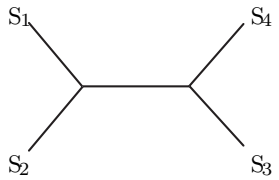


Figure 3.3. A 4-taxon tree.

For a second construction of invariants, again consider the 4-state general Markov model on the tree of Figure 3.3, and imagine the root placed midway along the central edge. Then, conditioned on the state at the root, state changes on the right half of the tree are independent of those on the left half of the tree. If only one state were possible at the root, then a $4^2 \times 4^2$ table of the joint distribution entries (with one dimension referring to states at the leaves in the left of the tree and one to the right) would be a contingency table for two independent 16-state variables, and hence a rank 1 matrix. Since 4 states are possible at the root, this table is instead a rank 4 matrix. But this implies all $5 \times 5$ minors (subdeterminants) of this matrix will be 0. Thus each such minor gives a degree 5 invariant for the model, and collectively these $\binom{16}{5}^2$ polynomials encode conditional independence of evolution on the two halves of the tree.

Lake (1987), and many papers following shortly thereafter, focused solely on finding and using linear invariants, the degree 1 polynomials in $I_{T,\mathcal{M}}$. (Reasons for this will be discussed in Section 5.) Cavender and Felsenstein (1987) found higher degree invariants, but worked primarily with a 2-state version of a Jukes-Cantor model. In subsequent research, some of which is quite recent, phylogenetic invariants have been investigated thoroughly for the Jukes-Cantor and Kimura models, and for the general Markov model.

Although we defer a survey of these results until Section 6, we informally describe them here: For those models most thoroughly investigated, *invariants for an arbitrary tree can be understood through those associated to local structures in the tree.* More particularly, specific invariants can be associated to each edge in a tree, and to each internal node in a tree, and from these one can produce all, or at least 'most', invariants. Examples of such invariants have already appeared above: the invariants encoding conditional independence statements about evolution on two parts of a tree separated by an edge.

We note that the association of invariants with local features in a tree gives a further compelling motivation for their investigation. In practical applications, one can hope to exploit this feature of invariants to provide measures of data support for local features within a tree, and then to use such measures as a means of inferring a tree from data by identifying local features.

## 4. Theoretical Insights from Invariants

Despite being originally proposed for use in inference, to date invariants have been most useful in providing theoretical understanding.

If a model is to be used for inference, it is of course crucial that the parameters of interest be identifiable. For phylogenetic models, the most important of these parameters is certainly the tree. While identifiability of the tree topology was known for models such as Jukes-Cantor, Kimura, and even general Markov, for mixture models allowing rate variation across sites little was known until recently.

For non-mixture models, tree identifiability was approached as follows. First, a natural distance between leaves of a tree was defined, and computed from 2-marginals of the joint distribution. Then these distances could be used to identify the topology, using equation (3.1). Unfortunately, this approach is not applicable to mixture models, since there is no known means of defining computable distances in such a setting.

Allman and Rhodes (2006) used phylogenetic invariants to obtain the first tree identifiability results for general sorts of mixtures. A general-Markov-like model which allows more states at internal nodes of the tree than at the leaves is introduced in that paper. Then invariants expressing conditional independence statements on the internal edges of the tree are constructed. The vanishing of these invariants can then be shown to identify the tree topology for generic values of model parameters. Although this model is far too general to be useful in data analysis, many models of more direct biological interest can be embedded in it, yielding results like the following.

**Theorem 1.** (Allman and Rhodes (2006)). *Consider a binary tree and the* 4*-state general Markov model. Then for mixture models allowing up to* 3 *rate classes, the tree topology is identifiable for generic choices of parameters.*

Of particular interest is another type of rate variation model called the *covarion* model (Tuffley and Steel (1998)). This is a model that allows a character to switch between 'on' and 'off' states as it evolves over a tree, modeling the intermittent presence of functional constraints in living organisms. When 'on' a standard Markov model of state change applies; when 'off' no state changes can occur. The on-off changes are themselves governed by a Markov process. As this model can also be embedded inside one allowing more states at internal nodes than leaves, further arguments show the following.

**Theorem 2.** (Allman and Rhodes (2006)). *Under the covarion model on a binary tree, for generic choices of parameters the tree topology is identifiable.*

Note that the covarion model is formulated as a continuous-time model, and is not itself algebraic. Nonetheless, tree identifiability for it has only been established by algebraic arguments.

In another direction, invariants have been used to investigate issues of multiple maxima in maximum likelihood inference of trees and numerical parameters.

Maximum likelihood inference of phylogenetic trees is generally performed by using some approximate optimization scheme to explore the space of numerical parameters, combined with a heuristic search of tree space (unless the number of taxa is small enough for a complete tree search). Thus one seldom has a guarantee that a true optimum has been found. While data sets had been found for which standard software finds multiple local optima (Salter (2001)), theoretical understanding of the situation has lagged behind.

For small trees and simple models, one might try to determine the maximum likelihood point by equating the gradient of the log-likelihood function to zero and solving the resulting equations algebraically. This was done in a simple 3-taxon case by Yang (2000). Subsequently, 3- and 4-taxon cases were considered by Chor, Hendy, Holland and Penny (2000), Chor, Hendy and Penny (2001), Chor, Khetan and Snir (2003), Chor, Hendy and Snir (2006) and Chor and Snir (2004). However, in order to make the algebra tractable, these later papers typically formulate the maximization problem not in terms of the parameters of the model, but rather in terms of the unknown joint distribution. Thus a constrained optimization problem, with phylogenetic invariants providing the constraints, is posed. Using computational algebra software, this problem can then solved exactly. Among the results are examples of 4-taxon data for which a 2-state model similar to the Jukes-Cantor leads to the maximum of the likelihood function occurring for two different trees and a continuum of numerical parameters on each tree.

Hoşten, Khetan and Sturmfels (2005) present a more general investigation of algebraic approaches to likelihood maximization, in both constrained and unconstrained formulations. In one phylogenetic example, using data from 4 taxa and the Jukes-Cantor model, they determine algebraically that there are multiple local maxima. While a similar computation with a more general model or more taxa is not yet feasible, it is remarkable that any such likelihood calculation has been done exactly.

## 5. Steps Toward Practical Applications

One attractive feature of Lake's proposal to use linear invariants for phylogenetic inference (Lake (1987)) is that it permits a weakening of the i.i.d. assumption that all sites evolved according to the same process. Specifically, suppose all sites are described by the same model $\mathcal{M}$, but perhaps with different choices of numerical parameters. Then such a rate-variation mixture model gives a joint distribution on which any linear invariant of $\mathcal{M}$ will vanish. We can see this geometrically: The zero-set of a linear invariant is a hyperplane. If we have a

collection of joint distributions on such a hyperplane, then any weighted sum of them lies in the hyperplane as well. Thus any phylogenetic inference scheme based solely on linear invariants will be insensitive to this type of rate variation.

Unfortunately, Lake's linear invariant method was not found to be very useful for practical inference. Although the method is statistically consistent, several works (most notably, Huelsenbeck (1995)) used simulation studies to show that it was extremely inefficient compared to other methods. Even when data was simulated from the assumed model, the sequence length needed to reliably infer the generating tree was very much greater for Lake's method than, for instance, maximum likelihood.

From the inefficiency of Lake's linear invariant method a perception has grown that any practical use of invariants will require much longer sequences than other methods to perform reliably, even in the absence of rate-variation mixtures. However, both theory and recent simulation studies show there is no basis for this view.

Consider, for instance, the application of Lake's invariants to data generated from a Kimura 3-parameter model on a tree relating 4 taxa. Then testing for the vanishing of Lake's two linear polynomials on an observed joint distribution amounts to testing whether that joint distribution lies on (or near) two particular hyperplanes (i.e., a linear space of dimension 61) in a space of dimension $4^4 - 1 = 63$. However, the set of all joint distributions arising from this model is much smaller than this. There are only 15 numerical parameters for the model on this tree, and one can show the phylogenetic variety has dimension exactly 15. There are many more phylogenetic invariants for this model, of degree greater than 1, which are needed to test whether a point is on or near the 15-dimensional set. It is because the variety is 'curved' and not 'flat' that these are non-linear, and ignoring the non-linear invariants simply throws away too much of the information in the data. If one considered *all* invariants for the Kimura model, one could potentially have a much more efficient method.

While this argument helps explain why hopes for Lake's invariants were not realized, there are further reasons why one should suspect all invariants will perform better. As was pointed out in Section 3, the equality of the 4-point condition can be expressed as a higher-degree invariant. But the neighbor joining algorithm (Saitou and Nei (1987)) is built on combining the 4-point condition for distances with simple averaging ideas, and so one might expect a clever use of invariants to at least match neighbor joining in performance. (Although neighbor joining can be justifiably criticized for its lack of a firm statistical framework, because of its speed and reasonably good performance it remains an important heuristic.)

Theoretical arguments aside, the same sort of simulation studies that showed Lake's invariants to be inefficient have recently shown that uses of non-linear

invariants can be highly efficient. Casanellas, Garcia and Sullivant (2005) and Casanellas and Fernández-Sánchez (2007) investigate the performance of one scheme using invariants to infer trees from data generated under a Kimura 2-parameter model. First a particular generating set of 8,002 polynomials $f_{i,j}(p)$ was chosen for $I_{T_j,\mathcal{M}}$, where $T_j$, $j = 1, 2, 3$, denotes one of the three topologically-distinct unrooted trees that might relate 4 taxa, and $\mathcal{M}$ was the Kimura 3-parameter model. The near-vanishing of all phylogenetic invariants for $T_j$ on an observed joint distribution $\widehat{P}$ was measured by the statistic

$$s_j = \sum_i \left| f_{i,j}(\widehat{P}) \right|,$$

and the tree $T_j$ was chosen if $s_j$ was the smallest of the three such statistics. Although the results in the first of these papers were promising, in that the correct tree was inferred a high percentage of the time even without long sequences, a lack of comparison to other methods made interpreting the results difficult.

Casanellas and Fernández-Sánchez (2007) remedy this by following the methodology of Huelsenbeck (1995), so that easy comparisons could be made to other inference methods. For that part of parameter space for 4-leaf trees that was considered, this invariant-based method is indeed quite efficient, with performance comparable to that of maximum likelihood. Furthermore, in some small regions of parameter space, it appears that invariants may even be more efficient than any of the methods considered by Huelsenbeck (1995).

Note that the statistic above is dependent on the precise choice of generators of the phylogenetic ideal. Although the generators used in these works were natural ones from an algebraic point of view, it is unclear whether others might behave better in inference; we do not yet know whether the invariants that are most useful statistically will be the ones that are algebraically simplest.

Also, these works were limited to inference of 4-taxon trees. As the number of invariants needed to generate the full ideal grows considerably as the number of taxa increases, it seems unlikely that a straightforward extension of the above statistic will ultimately prove useful. Whether all 4-taxon invariants can lead to a useful quartet-method for inference of larger trees, or other sets of invariants associated to the local structure of the tree can be effectively exploited, remains to be investigated. Nonetheless these studies have clearly shown that invariants have potential application in data analysis, and are worthy of further investigation.

Another method of inferring a tree was developed by Eriksson (2005) using ideas based on invariants, even though the invariants themselves do not appear in the final implementation. For the general Markov model of DNA evolution on an arbitrary tree, the invariants associated to an edge in the tree (or, equivalently, to the bipartition of taxa produced by deleting that edge), arise from the fact

that a certain matrix constructed from the joint distribution must have rank at most 4, as was indicated in Section 3. If one considers a matrix constructed from an observed distribution in this way for some bipartition of the taxa, then it should be close to a rank 4 matrix if there is an edge in the true tree inducing that bipartition. Thus an algorithm based on edge invariants might simply test which matrices associated to bipartitions are nearly of rank 4. As the singular value decomposition provides a good means of calculating how far a matrix is from having any given rank, and numerical algorithms for computing the SVD are highly developed, it is convenient to replace use of the edge invariants with SVD considerations.

The algorithm of Eriksson (2005) proceeds as follows, in a way reminiscent of neighbor joining: First observe that any tree relating four or more taxa must have at least two cherries (a cherry is a pair of taxa each joined to a common vertex by a single edge). Then consider all possible bipartitions of taxa in which one set has only two elements (a potential cherry). Use the SVD to choose the bipartition for which the corresponding matrix is closest to rank 4. Then join the two taxa together in a cherry. Treating this cherry as if it were a single taxon, we now have one fewer taxon and can repeat the process until all taxa are joined.

Although this algorithm is quite fast and performs reasonably well, in simulation studies it does not perform as well as maximum likelihood or even neighbor joining. (To some degree the studies were biased against the SVD method, which assumes only a general Markov model, while the other methods were performed assuming a more restrictive model in line with that used to simulate the data.) While the poor performance compared to neighbor joining is disappointing, the novel introduction of the SVD as a means of measuring the presence of a bipartition of the taxa is intriguing.

In investigating the performance of the Eriksson SVD algorithm on simulated data, we found it had a tendency to infer trees with too many cherries. While there are several factors that contribute to this, one of the most important ones has an interesting basis in geometry. If $n$ taxa are to be related, then the various matrices one must consider to test for the presence of bipartitions of $k$ and $n-k$ taxa are of size $4^k \times 4^{n-k}$, where $2 \leq k \leq n/2$. Within the $4^n$-dimensional space of $4^k \times 4^{n-k}$ matrices, however, the variety of rank 4 matrices can be shown to have dimension $4^{k+1} + 4^{n-k+1} - 16$. In particular, the dimension of the variety of rank 4 matrices is largest when $k$ is small. It is therefore easier for a 'random' matrix to be nearly of rank 4 when $k = 2$ than when $k$ is larger. As a result, the Eriksson algorithm tends to form too many cherries before it chooses to join cherries into larger groups. This issue of varying dimensionality of models is of course common in statistical comparisons, and finding an effective correction in this setting will be interesting work for the future.

It is of course easy to criticize these attempts to use invariants for tree inference as *ad hoc* and not sufficiently 'statistical'. In our view, there is much to be done to understand the performance and usefulness of invariants both within a statistical framework, and as potential contributors to heuristics that might aid in improving the performance of software implementations of more standard methods. Phylogenetic methods are full of compromises between biological reality and computational feasibility. Invariants are potentially useful in further developing methods precisely because they provide a different perspective.

## 6. Phylogenetic Ideals

We now survey in more detail the major results on determining phylogenetic invariants for specific models.

The Jukes-Cantor, Kimura 2-parameter, and Kimura 3-parameter models are the most biologically-plausible examples of *group-based* models. This means the character states can be identified with elements of an abelian group (the states $A, G, C, T$ are identified with elements of $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ for the Jukes-Cantor and Kimura models), and then probabilities of a state change from state $x$ to state $y$ along a particular edge are assumed to depend only on $y - x$. This special structure of group-based models allows one to use Fourier analysis in analyzing them. Although the mathematical ideas behind this are quite nice, in the end the key result is both simple and powerful.

**Theorem 3.** *For a group-based model, there are linear changes of variables for the parameter space of the model and for the joint distribution space, so that the parameterization map $\phi_{T,\mathcal{M}}$ in these new variables is given by monomial formulas.*

This monomial parameterization is the key to constructing invariants for the group-based models. This change of variables was first noticed for a 2-state group-based model (Hendy and Penny (1989)), and is sometimes referred to as a Hadamard transform. Evans and Speed (1993) exploited it to construct invariants for the Jukes-Cantor and Kimura models, as subsequently did Steel, Székely, Erdös, and Waddell (1993). The Fourier transform was generalized to all group-based models by Székely, Steel and Erdős (1993). Recently, however, Sturmfels and Sullivant (2005) pushed this line of reasoning forward to analyze the full ideal of phylogenetic invariants for these models. Varieties parameterized by monomial maps are known as *toric varieties* and are well-enough understood that additional theoretical understanding and computational tools could be applied.

**Theorem 4.**(Sturmfels and Sullivant (2005)). *For the Jukes-Cantor and Kimura models on any binary tree $T$, a set of generators of the phylogenetic ideal can be*

*explicitly given. Each of these generators is naturally associated to either an edge of the tree or a vertex.*

The generators associated to an edge express conditional independence statements, as discussed before. The generators associated to a vertex of a binary tree arise from those associated to an unrooted star tree with 3 leaves. If $T$ is not binary, then one only needs to understand the invariants for star trees with more than 3 leaves to extend the above theorem. For small numbers of leaves these can be determined computationally, but we still lack a theoretical analysis of them.

A helpful compendium of invariants for trees of up to 5 taxa, for group-based models of 2 and 4 states, is provided for easy downloading on the website created by Casanellas et al. (2005). This should be a useful resource for those wishing to explore the statistical behavior of these polynomials on data, simulated or real.

Invariants for the general Markov model were constructed by Allman and Rhodes (2003). Note that, unlike for group-based models, we lack a means of simplifying the parameterization map for the general Markov model. The key results of that paper actually focus on the 3-leaf star tree, and use rather simple observations on the structure of the model on that tree, together with basic linear algebra, to find explicit invariants. For the $k$-state model, these are of degree $k+1$, which is known to be the minimal degree possible. For $k \leq 4$, one can check these give all invariants of the minimal degree. Though for $k = 4$ the invariants have hundreds of terms, they can be naturally expressed as entries in a concise matrix formula involving classical constructions such as cofactors. While for 2- and 3-state models this construction gives a set of generators of the ideal for the 3-leaf tree, for the 4-state case relevant to DNA it is known that additional higher-degree invariants are needed (see Allman and Rhodes (2007a) for details). Nonetheless, one can also characterize the possible location of 'extraneous zeros' of this set of invariants as being in the zero set of another explicit polynomial. Thus while gaps remain in our knowledge of invariants for the 3-leaf case, substantial progress has been made.

Although trees relating more than three taxa are considered by Allman and Rhodes (2003), considerably more progress is made in determining all the invariants for the general Markov model on larger trees by Allman and Rhodes (2007a). In that work, actions of the matrix groups $GL_k$ on the phylogenetic varieties serve as an important tool. Representative results include the following.

**Theorem 5.**(Allman and Rhodes (2007a)) *Consider the $k$-state general Markov model on an arbitrary binary tree. If $k = 2$, then an explicit set of generators of the phylogenetic ideal can be given. If $k = 3$, then explicit polynomials can be given whose zero-set is the phylogenetic variety. For any $k$, if explicit polynomials are known whose zero set is the phylogenetic variety for the 3-leaf tree, then*

*explicit polynomials can be given whose zero set is the phylogenetic variety for the given tree.*

The first statement here establishes a conjecture of Pachter and Sturmfels (2004), that can be interpreted as saying that for $k = 2$ and binary trees, all phylogenetic invariants are generated by those associated to conditional independence statements on edges of the tree.

The results for higher $k$ are not quite as complete. First, we still lack complete knowledge of ideal generators for the 3-taxon tree if $k \geq 4$. Second, the proof that ideal generators can be given when $k = 2$ takes advantage of an 'accidental fact' that in that case there are no invariants for the 3-leaf tree. Without this fact for higher $k$, nothing is proved about ideal generators.

Finally we note that Casanellas and Sullivant (2005) consider invariants for the 4-state *strand symmetric model* for DNA. This model is a sort of amalgamation of a 2-state general Markov model with a 2-state group-based model, and so ideas arising for those constituent models can be applied to it to obtain detailed results on invariants. Lying between the group-based models and general Markov models, the strand symmetric model might, in some circumstances, offer more biological realism without excessive generality. (Bielawski and Gold (2002) provide one biological investigation of strand-symmetry.)

## 7. Further reading

Those seeking a broad overview of phylogenetics should consult Felsenstein (2004a). Semple and Steel (2003) also provide a good entry to the field, focusing on combinatorial aspects. The collections of articles edited by Gascuel (2005) and Gascuel and Steel (2007) provide a wide spread of perspectives, from the biological to mathematical, from theoretical to practical. In the second of these, Allman and Rhodes (2007b) give a more detailed exposition along the lines of this article, suitable for researchers in phylogenetics who are not algebraists. That work also includes a more extensive bibliography on phylogenetic invariants. Finally, Eriksson, Ranestad, Sturmfels and Sullivant (2004) introduce phylogenetics in a manner suitable for researchers in algebraic geometry.

## References

Allman, E. S. and Rhodes, J. A. (2003). Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.* **186**, 113-144.

Allman, E. S. and Rhodes, J. A. (2006). The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comput. Biol.* **13**, 1101-1113. arXiv:q-bio.PE/0511009.

Allman, E. S. and Rhodes, J. A. (2007a). Phylogenetic ideals and varieties for the general Markov model. *Adv. in Appl. Math.* to appear. arXiv:math.AG/0410604.

Allman, E. S. and Rhodes, J. A. (2007b). Phylogenetic invariants. In Gascuel, O. and Steel, M., editors, *New Mathematical Models of Evolution*. Oxford University Press.

Bielawski, J. P. and Gold, J. R. (2002). Mutation patterns of mitochondrial H- and L-strand DNA in closely related cyprinid fishes. *Genetics* **161**, 1589-1597.

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In *Mathematics in the Archeological and Historical Sciences*. Edinburgh University Press, Edinburgh.

Casanellas, M. and Fernández-Sánchez, J. (2007). Performance of a new invariants method on homogeneous and nonhomogeneous quartet trees. *Mol. Biol. Evol.* 24, 288-293. arXiv.org:q-bio.PE/0610030.

Casanellas, M., Garcia, L. D., and Sullivant, S. (2005). Catalog of small trees. In *Algebraic Statistics for Computational Biology* (Edited by L. Pachter and B. Sturmfels), 291-304. Cambridge University Press. http://www.math.tamu.edu/˜ lgp/small-trees/.

Casanellas, M. and Sullivant, S. (2005). The strand symmetric model. In *Algebraic Statistics for Computational Biology* (Edited by L. Pachter and B. Sturmfels), 305-321. Cambridge University Press.

Cavender, J. A. and Felsenstein, J. (1987). Invariants of phylogenies in a simple case with discrete states. *J. Classification* **4**, 57-71.

Chor, B., Hendy, M. and Penny, D. (2001). Analytic solutions for three-taxon $\mathrm{ML}_{MC}$ 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.

Chor, B., Hendy, M. and Snir, S. (2006). Maximum likelihood Jukes-Cantor triplets: Analytic solutions. *Mol. Biol. Evol.* 23(3), 626-632. `arXiv:q-bio.PE/0505054`.

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

Chor, B., Khetan, A. and Snir, S. (2003). Maximum likelihood on four taxa phylogenetic trees: Analytic solutions. *RECOMB'03*, 76-83.

Chor, B. and Snir, S. (2004). Molecular clock fork phylogenies: Closed form analytic maximum likelihood solutions. *Syst. Biol.* **53**, 963-967.

Eriksson, N. (2005). Tree construction using singular value decomposition. In *Algebraic Statistics for Computational Biology* (Edited by L. Pachter and B. Sturmfels), 347-358. Cambridge University Press.

Eriksson, N., Ranestad, K., Sturmfels, B. and Sullivant, S. (2004). Phylogenetic algebraic geometry. In *Projective Varieties with Unexpected Properties; Siena, Italy* pages 237-256, Berlin. de Gruyter. `arXiv:math.AG/0407033`.

Evans, S. N. and Speed, T. P. (1993). Invariants of some probability models used in phylogenetic inference. *Ann. Statist.* **21**, 355-377.

Felsenstein, J. (2004a). *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Felsenstein, J. (2004b). *Phylip: Phylogeny inference package*. Version 3.6. University of Washington.

Gascuel, O. (2005). *Mathematics of Evolution and Phylogeny*. Oxford University Press, Oxford.

Gascuel, O. and Steel, M. (2007). *New Mathematical Models of Evolution*. Oxford University Press, Oxford.

Hendy, M. D. and Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Systematic Zoology* **38**, 297-309.

Hoşten, S., Khetan, A. and Sturmfels, B. (2005). Solving the Likelihood Equations. *Found. Comput. Math.* **5**, 389-407. arXiv:math.ST/0408270.

Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Syst. Biol.* **44**, 17-48.

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

Pachter, L. and Sturmfels, B. (2004). Tropical geometry of statistical models. *Proc. Natl. Acad. Sci. USA* **101**, 16132-16137 (electronic).

Ronquist, F. and Huelsenbeck, J. P. (2003). MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574.

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

Salter, L. (2001). Complexity of the likelihood surface for a large DNA dataset. *Syst. Biol.* **56**, 970-978.

Semple, C. and Steel, M. (2003). *Phylogenetics* volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford.

Steel, M. (1994). Recovering a tree from the leaf colourations it generates under a Markov model. *Appl. Math. Lett.* **7**, 19-23.

Steel, M., Székely, L., Erdös, P. L., and Waddell, P. (1993). A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. *N. Z. J. Botany* **31**, 289-296.

Steel, M., Székely, L., and Hendy, M. (1994). Reconstructing trees from sequences whose sites evolve at variable rates. *J. Comput. Biol.* **1**, 153-163.

Stefankovic, D. and Vigoda, E. (2007). Phylogeny of mixture models: Robustness of maximum likelihood and non-identifiable distributions. *J. Comput. Biol.* **14**, 156-189. arXiv:q-bio.PE/0609038.

Sturmfels, B. and Sullivant, S. (2005). Toric ideals of phylogenetic invariants. *J. Comput. Biol.* **12**, 204-228. arXiv:q-bio.PE/0402015.

Swofford, D. (2002). *PAUP*: Phylogenetic analysis using parsimony* (** and other methods*). Version 4.0. Sinauer Associates.

Székely, L. A., Steel, M. A., and Erdős, P. L. (1993). Fourier calculus on evolutionary trees. *Adv. Appl. Math.* **14**, 200-210.

Tuffley, C. and Steel, M. (1998). Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* **147**, 63-91.

Yang, Z. (2000). Complexity of the simplest phylogenetic estimation problem. *Proc. R. Soc. Lond. Ser. B* **267**, 109-116.

Department of Mathematics and Statistics, University of Alaska Fairbanks, PO Box 756660, Fairbanks, AK 99775, U.S.A.

E-mail: e.allman@uaf.edu

Department of Mathematics and Statistics, University of Alaska Fairbanks, PO Box 756660, Fairbanks, AK 99775, U.S.A.

E-mail: j.rhodes@uaf.edu

(Received October 2006; accepted April 2007)