4

PHYLOGENETIC INVARIANTS

Elizabeth S. Allman and John A. Rhodes

Abstract

Under many common models of sequence evolution along trees, frequencies of base patterns in extant taxa satisfy certain polynomial relationships known as 'phylogenetic invariants'. Though introduced in 1987 for phylogenetic inference, invariants remained difficult to construct, and the inefficiency of simple inference schemes based on known linear ones was discouraging. Recently there has been much progress in producing phylogenetic invariants, and in understanding their structure. Potentially useful connections between specific topological features in a tree (vertices and nodes) and specific invariants have emerged. We introduce some of the mathematical ideas underlying current understanding of invariants, with an emphasis on a geometric viewpoint and rank computations. We also highlight new insights arising from invariants, including better understanding of maximum-likelihood estimation and proofs of the identifiability of certain substitution models, such as the covarion and mixture models.

4.1 Introduction

Probabilistic models for the evolution of biological sequences are used throughout phylogenetics, both for theoretical analysis and for practical inference. Basic assumptions in these models lead naturally to expressing their predictions through *polynomial* expressions. This simple observation leads to the insight that polynomial algebra can provide alternative perspectives in phylogenetics.

Phylogenetic invariants were introduced in 1987 in two independent works, by Cavender and Felsenstein [13], and by Lake [48]. For DNA sequences, phylogenetic invariants are polynomial relationships that must hold between the frequencies of various base patterns in idealized data, which is perfectly in accord with a particular model and tree. By testing whether such polynomials for various trees were 'nearly zero' when evaluated on the observed frequencies of patterns in real data sequences, it was hoped that one could infer which tree best explained the data.

CUEL: "CHAP04" — 2007/4/3 — 18:06 — PAGE 108 — #1

A number of difficulties, which will be surveyed later in this chapter, prevented invariants from being quickly developed into useful inference tools. In particular, while Lake's linear invariants had some desirable statistical properties, practical inference based on them performed poorly on sequences of a length typical of real data. This perhaps led some to question the value of invariants in general, even though few serious attempts at using higher degree invariants for inference were made. Indeed, thorough knowledge of non-linear invariants was largely lacking for DNA models, with the notable exception of the results on group-based models that began with Evans and Speed [24]. As the need for more general models to adequately describe data had become clearer, invariants that incorporated the added complexity were simply not known.

Recently, however, much progress has occurred in understanding phylogenetic models algebraically. Our knowledge of phylogenetic invariants has grown to include models of sufficient generality to encompass some of those currently used for inference. Most importantly, for those models which are well understood, a close relationship holds between specific invariants and particular local topological features of trees, such as edges or nodes. While more remains to be done in determining the structure of invariants for additional models of interest, there is now enough understanding to consider again how we might use invariants, either for inference or for theoretical analysis.

This chapter is divided into two parts. In the first, we discuss constructions of invariants, explain how they can be interpreted, and survey results on the extent to which all invariants for various models are known. We begin with a careful development of some invariants for the general Markov model, in order both to be concrete and to emphasize that invariants make interpretable statements about statistical models. In the second part, we turn to applications of invariants. These include recent investigations focused on understanding when maximum-likelihood inference may face multiple local optima, and on establishing the identifiability of tree topologies for certain mixture models. We end with more speculative uses for practical inference. We hope to convey that the perspectives invariants offer on phylogenetic models can be valuable in many settings, and that more applications remain to be discovered.

The mathematical field most appropriate to studying phylogenetic invariants is *algebraic geometry*, which is rich and well-developed, but far from the typical background of most phylogenetics researchers. In this chapter we provide only a gentle introduction to its terminology when necessary, and our presentation of some results omits more technical details. We hope this creates an overview that will be especially useful for those who might be more interested in thinking about how to use invariants than in how to find them.

A first example. For a concrete introduction to viewing a probabilistic model in phylogenetics algebraically, consider the following illustrative example:

An ancestral sequence at the root r of a tree gives rise to two descendant sequences, at leaves a and b of the tree shown in Fig. 4.1. We model evolution at a single site in a sequence, with the idea that each site evolves according to the same model, but independently (the i.i.d. assumption).

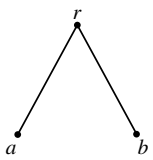

FIG. 4.1. Two taxa a and b descend from a common ancestor r.

For the ancestral sequence at r we specify the probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$ with which the four bases ($A = 1$, $G = 2$, $C = 3$, $T = 4$) might appear at a particular site, or equivalently by the i.i.d. assumption, the relative frequencies at which bases appear across all sites. For each edge of the tree, we model the evolutionary process by specifying probabilities of various substitutions occurring. Thus for edge e_1, leading from r to a, we specify a 4×4 matrix M_1 whose (i, j)-entry is the conditional probability of observing base j in the sequence at a given that the ancestral base at r was i. Similarly, a matrix M_2 describes the mutation process on edge e_2, leading from r to b. The parameters of the model, the 4-*state general Markov model*, are the tree of Fig. 4.1 along which we model evolution, and the entries of $\boldsymbol{\pi}, M_1$, and M_2.

From the model parameters we compute the probability of each possible observation. The probability of seeing base j in a site at a and base k in the same site at b is

$$\text{Prob}(a = j,\ b = k) = p_{jk} = \sum_{i=1}^{4} \pi_i M_1(i, j) M_2(i, k). \tag{4.1}$$

The *joint distribution* of bases $P = (p_{jk})$, then, can be thought of as a 4×4 matrix, each of whose entries is a 4-term degree-3 polynomial in the parameters of the model. These 16 polynomials parameterizing the model reflect all the modeling assumptions, including the substitution probabilities, and the tree topology of Fig. 4.1.

In order to produce a clear and instructive example, we simplify the model further (at the expense of biological plausibility) by restricting it to the situation where the ancestral sequence is composed of only the base A, so that $\boldsymbol{\pi} = (1, 0, 0, 0)$. This *ancestral-A model* leads to a simplification of equation (4.1) so that the joint distribution of bases is given by the 16 quadratic polynomials

$$p_{jk} = M_1(1, j) M_2(1, k). \tag{4.2}$$

Now from inspecting equation (4.2), we observe that

$$p_{jk} p_{mn} - p_{jn} p_{mk} = 0, \tag{4.3}$$

since each term in this difference can be expressed in terms of the parameters as

$$M_1(1, j) M_2(1, k) M_1(1, m) M_2(1, n).$$

Thus for every choice of j, k and m, n we have found a polynomial,

$$f_{jk,mn}(P) = p_{jk}p_{mn} - p_{jn}p_{mk},$$

that will evaluate to 0 when $P = (p_{jk})$ is any true distribution of bases arising from the ancestral A-model, without regard to the particular numerical values appearing in the Markov matrix parameters. These polynomials are called *invariants* for the ancestral-A model on a 2-taxon tree.[1]

More generally, an invariant for a model is a polynomial that gives zero when evaluated on any distribution arising from that model, regardless of the parameter values leading to that distribution. On a distribution that does not arise from the model, an invariant typically evaluates to give a non-zero result. Since the invariants found here will, in fact, vanish on distributions arising from ancestral-G, ancestral-C, and ancestral-T models also, they are better termed as invariants for an ancestral-1-base model. Even so, by allowing two or more ancestral states it is easy to construct numerical examples of distributions on which these invariants will not be zero.

To see why model invariants might be useful, imagine aligned DNA sequences from taxa a and b. We wish to test whether this data might have been produced from the ancestral-A model on the tree above. We record the observed distribution $\widehat{P}$, a 4×4 array giving frequencies of aligned bases in the two sequences. If we believe the model provides a good description of the data, then we suspect $\widehat{P} \approx P$, where P is a true distribution arising from the model for some unknown choice of parameters. Thus for any model invariant, f, since $f(P) = 0$, we should find that $f(\widehat{P}) \approx 0$.

Thus we might simply evaluate the model's invariants on the observed distribution $\widehat{P}$ and, if we get values close to zero, take that as evidence that the ancestral-A model might describe the data well. If we get values far from zero, we could take that as evidence against the ancestral-A model providing a good fit to the data.

This is schematically indicated in Fig. 4.2, where we imagine two alternative models leading to different sets of invariants. In order to choose which model may best describe a data point $\widehat{P}$, we wish to determine if $\widehat{P}$ is closer to the zero set of one collection of invariants or the other.

In this way polynomial invariants for more elaborate phylogenetic models might provide a method of inference that *circumvents determination of numerical parameters*. In particular, the tree topology may be of more intrinsic interest than the numerical parameters in a phylogenetic model. If invariants can be found that test for each possible tree topology for a set of taxa, evaluating them on an observed distribution to see if they nearly vanish might enable us to infer the topology.

[1] This model is actually a familiar one in statistics, outside of phylogenetics; it is the independence model for a 2-way table P. The invariants above are commonly expressed in a slightly different form, using an *odds ratio*: $\frac{p_{jk}p_{mn}}{p_{jn}p_{mk}} = 1$.

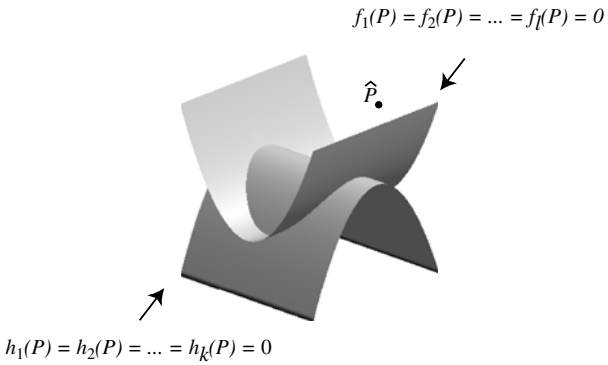

FIG. 4.2. The f_i and h_i are invariants for two alternative models. All joint distributions arising from the first model lie in the 'surface' defined by $f_i(P) = 0$, and similarly for the second. To decide which model better explains a data point $\widehat{P}$, we attempt to judge whether $f_i(\widehat{P}) \approx 0$ or $h_i(\widehat{P}) \approx 0$.

This idea, focused on determining the tree topology from larger sets of sequences, was the one introduced in both [13] and [48]. There are difficulties in applying this idea as naively as described here; nonetheless, it is a good one to keep in mind for motivation. In a nutshell, invariants have the potential to tell us something about whether an observed distribution might have arisen from a particular model, without having any need to infer numerical parameters.

Notice that in this example there are two sets of polynomials. The first, appearing in equation (4.2), are the *parameterization polynomials*, expressing the true distribution our model predicts in terms of the model parameters. The second, the invariants of the model, appearing in equation (4.3), describe the relationships that must hold within a distribution resulting from the parameterization. The parameterization polynomials are straightforward to produce, since they express the model as we have designed it. The invariants are consequences of the parameterization polynomials, but how to produce them or interpret their meaning is much less obvious for most models.

Finally, note that the idea of invariants need not be limited to phylogenetic models. Indeed, they can be studied in other statistical settings where polynomial parameterizations arise. The complexity and structure of phylogenetic models, however, makes the subject particularly rich in this setting.

Part 1. Finding Invariants

Discussing constructions of invariants first requires a more detailed specification of some phylogenetic models. Before proceeding, however, we note there is one invariant whose existence is easy to explain.

Consider any probabilistic model which allows only finitely many outcomes. The distribution will take the form of an array, where each entry is the probability

of one possible outcome. For instance, for DNA substitution models for n-taxa, the joint distribution can be given by an n-dimensional $4 \times \cdots \times 4$ array. The vanishing of the *stochastic invariant*,

$$\sum_{i_1,i_2,\ldots,i_n} p_{i_1 i_2 \ldots i_n} - 1,$$

where the summation is over all entries of the distribution, states that the probabilities of all possible outcomes must add to 1. It is therefore an invariant for every such model.

4.2 Phylogenetic models on a tree

For convenience, we will assume all trees are binary (i.e. trivalent at all internal nodes, except possibly bivalent at a root).

Let T be an n-leaf unrooted binary tree, with its leaves labeled by a collection of taxa $X = \{a_1, a_2, \ldots, a_n\}$. We may introduce a root r by either choosing some existing node of T, or subdividing some edge of T and choosing the new node as the root, obtaining the rooted tree T^r. In a rooted tree T^r we view all edges as directed away from r.

The *κ-state general Markov (GM) model* on T^r is a model of character evolution parameterized by:

1. A root distribution vector $\boldsymbol{\pi}_r = (\pi_1, \pi_2, \ldots, \pi_\kappa)$. We interpret π_i as the probability that the character is in state i in the ancestral taxon r. Thus $\pi_i \geq 0$ and $\sum_{i=1}^{\kappa} \pi_i = 1$. For simple DNA models, $\kappa = 4$.
2. For each directed edge e of the rooted tree, a $\kappa \times \kappa$ Markov matrix M_e. We interpret the (i,j)-entry of M_e as giving the conditional probability that the character is in state j at the descendant end of e given that it was in state i at the ancestral end. Thus $M_e(i,j) \geq 0$ and $\sum_{j=1}^{\kappa} M_e(i,j) = 1$.

A key feature of the model is that we may observe states only at the leaves of the tree; states at all internal nodes are *hidden*.

Rather than give a general formula for the joint distribution arising from this model, we indicate its form through an example, with a specific tree. Considering the tree of Fig. 4.3, with M_i denoting the Markov matrix for edge e_i, we find the entries of the joint distribution P are

$$P(i,j,k,l,m) = p_{ijklm} = \sum_{s=1}^{\kappa}\sum_{t=1}^{\kappa}\sum_{u=1}^{\kappa}\sum_{v=1}^{\kappa}[\pi_s M_1(s,i)M_2(s,t)M_3(t,u)\times M_4(u,j)M_5(u,k)M_6(t,v)M_7(v,l)M_8(v,m)]. \quad (4.4)$$

Note that these κ^5 parameterization polynomials in $8\kappa(\kappa-1)+\kappa-1$ variables reflect not only the assumptions of the general Markov model, but also the form of the tree in Fig. 4.3. Indeed, from the parameterization one can even reconstruct the tree, as it algebraically encodes the topology.

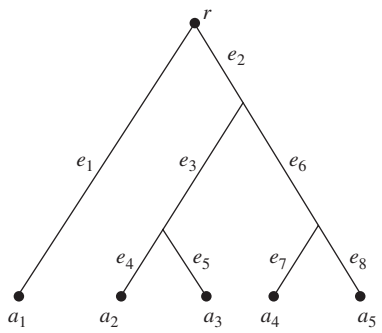

FIG. 4.3. A 5-taxon tree.

Most of the other models we consider are submodels of GM, in that they merely place additional restrictions on the form of the numerical parameters. The 2-*state symmetric model*, or Cavender–Farris–Neyman model, assumes $\kappa = 2$, $\boldsymbol{\pi} = (.5, .5)$, and that every Markov matrix has the form

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

where a_e is a scalar parameter. This is an example of a *group-based* model (see [60] for an explanation of this terminology). Note that with this assumption, the overall polynomial form of equation (4.4) is retained, but the degree of each term drops by one, and the polynomial involves only the variables a_e, for each edge e.

Other group-based models of particular interest include the Kimura 3-parameter model, a 4-state model that assumes $\boldsymbol{\pi} = (.25, .25, 25, .25)$ and

$$M_e = \begin{pmatrix} d_e & a_e & b_e & c_e \\ a_e & d_e & c_e & b_e \\ b_e & c_e & d_e & a_e \\ c_e & b_e & a_e & d_e \end{pmatrix},$$

where $d_e = 1 - a_e - b_e - c_e$. Specializing by requiring $b_e = c_e$ yields the Kimura 2-parameter model, and requiring $a_e = b_e = c_e$ yields the Jukes–Cantor (or 4-state symmetric) model.

It is common in other contexts to use phylogenetic models which have a continuous time formulation, where one specifies a rate matrix Q and edge lengths t_e to describe the substitution process, with the Markov matrix on an edge being $M_e = \exp(Qt_e)$. Usually the rate matrix for all edges is taken to be the same, which is a strong assumption of commonality about the substitution process over the entire tree. Indeed, typical implementations in software of the general time-reversible model (GTR) are of this sort. Note that such models do not have polynomial parameterizations, but rather ones involving exponentials.

In studying invariants, in order that the parameterization maps be polynomial, we do *not* assume a continuous-time model of base substitution, nor commonality of rates across the tree. Rather we use a discrete notion of time, in which the full evolutionary process on an edge of the tree is lumped together to be described by a single matrix. As a result, the models dealt with when studying or utilizing invariants are, in this respect, more general than those used in most software.

One might view the generality of GM as either a strength (if one doubts that the assumptions of a model such as the GTR are justified for a data set) or a weakness (if one believes those assumptions are justified, and extra generality in the model leads to the possibility of overfitting data). Regardless, note that a model such as the GTR is a submodel of GM, in that it merely places additional (non-algebraic) restrictions on the form of allowable parameters. Thus, whatever invariants allow us to say about the GM model will imply statements about its submodels such as GTR. In Section 4.9, for example, we describe an application of invariants to some continuous-time models.

We also note that the GM model does not allow any 'rate variation' across sites, so a model such as GTR+I+Γ is not a submodel. Later, in Section 4.9, we return to a discussion of rate variation, explaining in more detail how invariants can be used to understand both rate-matrix models with variation in rates across sites, and also the covarion model.

4.3 Edge invariants and matrix rank

An invariant f for the GM model on the particular tree T^r of Fig. 4.3 is a polynomial in κ^5 variables, the indeterminate entries p_{ijklm} of a $\kappa \times \kappa \times \kappa \times \kappa \times \kappa$ array P. Furthermore, when P is given numerical values P_0 produced by some choice of parameter values in equations (4.4), we have $f(P_0) = 0$. Even a glance at equations (4.4), however, indicates we have little chance of finding any invariants by the 'inspection' approach we used for the ancestral-A model of Section 4.1.

To construct a first class of invariants for this model, we proceed by building on the example of the Introduction. We again consider the much simpler situation of Fig. 4.1 and equation (4.1), in order to rederive its invariants in a more sophisticated way. Notice first that the 16 versions of equation (4.1) can be combined into a single matrix equation

$$P = M_1^T \operatorname{diag}(\boldsymbol{\pi}) M_2, \tag{4.5}$$

where $\operatorname{diag}(\boldsymbol{\pi})$ denotes a matrix with the vector $\boldsymbol{\pi}$ placed along the diagonal and with 0 in all off-diagonal entries.

For the ancestral-A model, we make the additional assumption that $\boldsymbol{\pi} = (1, 0, 0, 0)$, so that $\operatorname{diag}(\boldsymbol{\pi})$ has only one non-zero entry. With this assumption, then, equation (4.5) implies that the matrix P must have rank at most 1, for $\operatorname{diag}(\boldsymbol{\pi})$ is a matrix of rank 1, and the rank of a product of matrices is at most the minimal rank of the factors. But from linear algebra there is a well-known algebraic condition on the entries of a matrix of rank 1: A matrix has rank 1 if,

and only if, its 2×2 minors (determinants of submatrices chosen by picking 2 rows and 2 columns) are all zero. Since these minors are precisely the polynomials of equation (4.3), we have recovered our previous invariants for the ancestral-A model on a 2-taxon tree.

To develop this viewpoint further, we consider an ancestral-AG model on the same tree; that is, we assume the GM model with $\boldsymbol{\pi} = (\pi_A, 1 - \pi_A, 0, 0)$. Now since $\mathrm{diag}(\boldsymbol{\pi})$ has rank 2, again using that the rank of a product is at most the minimal rank of its factors, equation (4.5) establishes that the rank of P is at most 2. Thus all 3×3 minors of P give invariants. For an ancestral-AGC model, similar reasoning shows P has rank at most 3, and so $\det(P) = 0$ is the sole invariant we obtain.

For the 4-state GM model on the 2-leaf tree, where we place no restrictions on $\boldsymbol{\pi}$, we similarly conclude that P must have rank at most 4. However, since P is 4×4, there is no real content in this observation, since the rank of any matrix is bounded by its dimensions. Thus we obtain no invariants from this viewpoint (and indeed none exist for this model on the 2-taxon tree, except the stochastic one.)

To summarize the viewpoint so far, and ultimately to obtain invariants for more taxa, it will be helpful to consider a slight broadening of the model. We step beyond the phylogenetic setting, but still base our model on the graphical depiction of Fig. 4.1. We imagine 3 discrete random variables, associated to the nodes r, a, b. The variable at r may take on any of κ states, while those at a and b may take on any of λ and μ states, respectively. A κ element root distribution vector $\boldsymbol{\pi}$ specifies probabilities of states at r, while $\kappa \times \lambda$ and $\kappa \times \mu$ Markov matrices give transition probabilities to the various states at a and b. Finally, we observe states only at a and b, with those at r being hidden.

Under this model, we see that equation (4.5) still applies to give the joint distribution. We also see that since the diagonal matrix has rank at most κ, P will also have rank at most κ, and thus all $(\kappa + 1) \times (\kappa + 1)$ minors of P must vanish. Provided $\lambda, \mu > \kappa$, so that P is big enough for such minors to exist, we have found some invariants of the model.

These invariants, which test for matrix rank, have a direct statistical interpretation: They express the basic assumption of this model, that the stochastic processes occurring along the two edges leading from r are *independent, conditioned on the state at* r.

To use this observation to find invariants for the κ-state GM model, we must consider a tree with more taxa, such as that to the left in Fig. 4.4. Suppose the root r is located at the left of the internal edge. Then the GM model has as parameters a root distribution vector, and five $\kappa \times \kappa$ Markov matrices.

We can ignore some of the structure in the model by grouping together taxa, letting $a = \{a_1, a_2\}$ and $b = \{a_3, a_4\}$. The random variable associated to a now has κ^2 states, the pairs of states for a_1 and a_2, and similarly for b. The graphical depiction of the model is now that of the right side of Fig. 4.4, which is identical to Fig. 4.1. For this 'coarsened' model we can express the $\kappa \times \kappa^2$ matrix

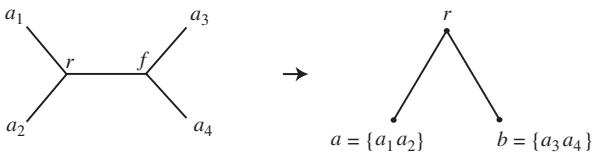

FIG. 4.4. A 4-taxon tree, with taxa a_1, a_2, a_3, a_4, rooted at r, and its coarsening to a simpler model.

parameters M_1 and M_2 in terms of the GM parameters:

$$M_1(i,(j,k)) = M_{ra_1}(i,j)M_{ra_2}(i,k),$$

$$M_2(i,(j,k)) = \sum_{l=1}^{\kappa} M_{rf}(i,l)M_{fa_3}(l,j)M_{fa_4}(l,k).$$

Coarsening the GM model in this way corresponds to changing the way we view the joint distribution array P. Though initially we viewed P as a $\kappa \times \kappa \times \kappa \times \kappa$ array, we now 'flatten' it to a $\kappa^2 \times \kappa^2$ matrix

$$\mathrm{Flat}(P)((i,j),(k,l)) = P(i,j,k,l).$$

Note that we have merely rearranged the way we view entries of P; the entries themselves are unchanged.

This coarsened GM is now an instance of a model for which we have already found invariants. We can therefore immediately see that all $(\kappa+1) \times (\kappa+1)$ minors of $\mathrm{Flat}(P)$ are invariants of the GM model on this tree, since the flattening of P must have rank at most κ. These invariants should be interpreted as expressing a conditional independence statement that the state-change process on the branches leading from r to a_1 and a_2 is independent of that on the edges leading from r to a_3 and a_4, conditioned on the state at r.

Despite appearances, these invariants do not actually depend on the location of r at one end of the internal edge of the tree. It can be shown that for a dense subset of all parameters, the GM model with one specified root location on a tree T produces the same joint distributions as the GM model with a different root location on T. This means we can freely move the root to a location convenient for our construction.

Note that the arrangement of entries in $\mathrm{Flat}(P)$, and thus the invariants we have found, depend only on the split of taxa $\{a_1,a_2\},\{a_3,a_4\}$ induced by the internal edge of the tree. We thus refer to these as *edge invariants* associated to the single internal edge of the tree.

This construction easily generalizes to larger trees. We can pick any internal edge of T and flatten P according to the resulting split. For a concrete example, consider the 2-state GM model on the 5-taxon tree of Fig. 4.5. Denoting states by 0 and 1, from the $2 \times 2 \times 2 \times 2 \times 2$ joint-distribution array P, we obtain two

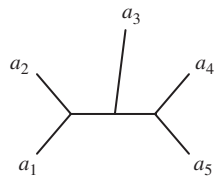

FIG. 4.5. A 5-taxon tree.

edge flattenings. The $\{a_1, a_2\}, \{a_3, a_4, a_5\}$ split gives

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

and the $\{a_1, a_2, a_3\}, \{a_4, a_5\}$ split gives

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

By what we have seen, all 3×3 minors of each of these matrices are invariants of the GM model on this particular tree.

4.4 Vertex invariants and tensor rank

The edge invariants of the GM model that are described in the last section express a conditional independence statement: character state-changes in the two parts of a tree separated by an edge are independent of one another, conditioned on the state of the character at some point along the edge.

Other invariants for the GM model express a similar sort of conditional independence statement, but focus on an internal node of the tree rather than an edge. To explain them, we first focus on the simplest tree for which they can arise, the 3-taxon tree with only one internal node, as in Fig. 4.6.

Here we imagine the central node is the root. Numerical parameters for the model then are the root distribution $\boldsymbol{\pi}_r$ and three $\kappa \times \kappa$ Markov matrices M_1, M_2, and M_3 giving probabilities of changes in state along the three edges leading from the root. The joint distribution for this model is a $\kappa \times \kappa \times \kappa$ array $P = (p_{ijk})$,

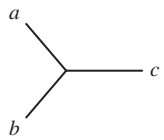

FIG. 4.6. The 3-taxon tree.

where

$$p_{ijk} = \sum_{l=1}^{\kappa} \pi_l M_1(l,i) M_2(l,j) M_3(l,k). \tag{4.6}$$

Since the matrix notation used in equation (4.5) is insufficient for describing a 3-dimensional array, we take an alternate approach. We first introduce arrays representing intermediate steps in equation (4.6): for each state l at the internal node, let P_l be the $\kappa \times \kappa \times \kappa$ array with ijk-entry $M_1(l,i)M_2(l,j)M_3(l,k)$. Notice that P_l is simply a joint distribution for an 'ancestral-base-l' model, similar to that of the introduction, but now for a 3-taxon tree.

The arrays P_l have a particularly simple structure, though. All entries are found by taking the various products of entries from the lth rows of M_1, M_2, and M_3. In other words, P_l is the *tensor product* of three rows. This parallels the situation for the 2-taxon tree in the last section, where the ancestral-A model had joint distribution $P = (p_{ij})$, with

$$p_{ij} = M_1(1,i)M_2(1,j)$$

so

$$P = \mathbf{r}_1^T \mathbf{r}_2,$$

where $\mathbf{r}_1$ was the first row of M_1 and $\mathbf{r}_2$ the first row of M_2. Just as this P was a rank 1 matrix, we call the 3-dimensional array P_l a *rank 1 tensor*. More formally, a 3-dimensional array is said to have rank 1 if it is the tensor product of 3 non-zero vectors.

When a 3-dimensional joint distribution is a rank 1 tensor, the fact that its entries are simple products of the form given here is just a manifestation of independence of the states for the 3 indices. Indeed, a rank 1 joint distribution occurs exactly when a model assumes a single state at the internal node of the graphical model of Fig. 4.6, with independent state changes on each edge leading away.

Now for the full model on the 3-taxon tree, we have that P is the weighted sum of κ rank 1 tensors,

$$P = \sum_{l=1}^{\kappa} \pi_l P_l,$$

with one summand for each of the κ possible states at the internal node. As the *tensor rank* of an array is the smallest number of rank 1 tensors needed to

express it as a sum, P is thus a tensor of rank at most κ since, just as before, it is a sum of rank 1 tensors. Emphasizing the statistical viewpoint, the joint distribution P is a tensor of rank at most κ precisely because of independence of the state changes on the edges of the tree, conditioned on the κ possible states at the root. This parallels the 2-taxon, matrix situation of the last section.

This gives a good way of thinking of invariants for the GM model on the 3-taxon tree: they should be interpreted as making a conditional independence statement about state changes on the 3 edges emerging from the internal node. But how do we explicitly find invariants for this model?

For edge invariants, we could use the classical results on the relationship between matrix rank and vanishing of minors. Although by general principles of algebraic geometry, analogs of matrix minors must exist for testing tensor rank, they are only explicitly known for tensors of a few special sizes.[2]

To find invariants for the 3-taxon tree, we need a direct construction, as supplied in [1]. While that paper gives a variety of invariants of different forms, the most important are the ones arising from *commutation relations* that are derived from an observation that certain expressions built from the joint distribution give commuting matrices. Even for the 4-state model, these invariants are rather complicated when expressed in ordinary polynomial notation; though each term is only of degree 5, there are hundreds of terms.

However, they can be given a concise expression using matrices. To illustrate a typical form, for any choice of state k let $P_{abk} = (p_{ijk})$ be the kth 'slice' of P, a matrix obtained by only considering those entries in the 3-dimensional array P with a fixed index of k in the 3rd position (corresponding to state k at taxon c). Then for any choice of i, j, k it can be shown that the matrix equations

$$P_{abk}\,\mathrm{Cof}(P_{abj})^T P_{abi} = P_{abi}\,\mathrm{Cof}(P_{abj})^T P_{abk} \tag{4.7}$$

must hold if P arises from the GM model. Here $\mathrm{Cof}(M)^T$ refers to the transpose of the co-factor matrix of M, which is a standard construction from linear algebra. As this equation expresses the equality of two $\kappa \times \kappa$ matrices, it gives κ^2 individual invariants from equating entries. Since each entry of the co-factor matrix is a polynomial of degree $\kappa - 1$, these invariants are of degree $\kappa + 1$.

When $\kappa = 2$, a calculation shows that all of these polynomials simply give 0. In fact, for the 2-state GM model on a 3-taxon tree, one can show the only invariant is the stochastic one, so this is as it should be.

For the 4-state model, however, one can verify that these polynomials are not zero. In fact, minor variations on the construction can produced 1728 linearly-independent degree 5-invariants. Other means [36, 49] can show this

[2]Tensor rank is a more subtle notion than one might expect from familiarity with the matrix concept. In particular, analogues of matrix minors will test for *border rank* rather than rank, since the closure of tensors of a certain rank may contain ones of higher rank. This phenomenon does not occur for matrices.

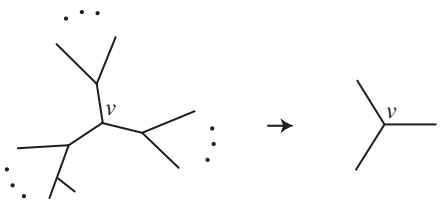

FIG. 4.7. Flattening a model at a vertex ν.

is the dimension of the full space of degree 5-invariants, and that except for the stochastic invariant there are essentially no others of lower degree.[3]

With some invariants in hand for the 3-taxon tree, a 'flattening' approach can be used again to give invariants for n-taxon binary trees. Picking any internal vertex v of the tree, we combine the taxa into three groups, as indicated in Fig. 4.7.

Coarsening our model in this way corresponds to rearranging the entries of the n-dimensional joint distribution array into a 3-dimensional array with size $\kappa^{n_1} \times \kappa^{n_2} \times \kappa^{n_3}$, where $n = n_1 + n_2 + n_3$. For a κ-state model, this flattened array must be a tensor of rank at most κ, since just as before it is a sum of rank 1 tensors, with one summand for each possible state at the internal node. Invariants for this coarsened model, which must also be invariants for the original model, are referred to as *vertex invariants*. With a bit of additional work [6], one can obtain explicit formulas for all vertex invariants provided one has them for the 3-taxon tree.

4.5 Algebraic geometry and computational algebra

Once invariants, such as the edge and vertex invariants for the GM model discussed above, have been found for a particular model, a natural question is whether there are others. To be able to discuss this properly, we informally introduce a little of the viewpoint and language of algebraic geometry.

Suppose we are given a collection of M polynomial functions, $g_1, g_2, \ldots, g_M$ depending on N variables $x_1, x_2, \ldots, x_N$. Allowing the variables to range over the complex numbers, we have a function

$$\phi : \mathbb{C}^N \longrightarrow \mathbb{C}^M,$$
$$(x_1, x_2, \ldots, x_N) \longmapsto (g_1(x_1, \ldots, x_N), \ldots, g_M(x_1, \ldots, x_N)).$$

We have in mind here that the g_i are the parameterization polynomials for the joint distribution of a phylogenetic model, the x_i are the parameters, and ϕ gives us the full joint distribution array for any parameter choice.

[3]While it is known that some additional invariants of degree 9 are also needed to obtain all invariants for the 3-taxon model, the full situation is not yet completely understood [6].

The image of ϕ, the set $\phi(\mathbb{C}^N)$, is a parameterized subset of $\mathbb{C}^M$, which we view as some sort of high-dimensional 'surface', which is smooth at most points, but perhaps has some singularities. For example, the cartoon depiction of Fig. 4.2 represents two such 'surfaces', though in practice dimensions are usually much higher.

We now try to describe $\phi(\mathbb{C}^N)$ *implicitly*, as the zero set of polynomials. Introducing variables $P = (p_1, p_2, \ldots, p_M)$, we look for polynomials in the p_i that vanish when $P = \phi(x_1, \ldots, x_N)$. Optimally, we determine the entire set $I = \{f\}$ of all polynomials f in the variables p_i, such that

$$P = (p_1, \ldots, p_M) \in \phi(\mathbb{C}^N) \text{ implies } f(p_1, \ldots, p_M) = 0.$$

Thus the vanishing of such an f on a point P would provide some evidence that it is in the image of ϕ.[4]

Indeed, this is precisely what we have been trying to do for phylogenetic models. In that context the set I of polynomials implicitly defining the set of joint distributions $\phi(\mathbb{C}^N)$ are the phylogenetic invariants.[5]

For phylogenetic models, or statistical models in general, the numerical parameters usually represent probabilities. It thus might seem more reasonable to require that parameters be in some subset of the interval $[0, 1]$, or at least in $\mathbb{R}$. However, in algebraic geometry it is well understood that many technical issues are easier dealt with when we allow variables to range over $\mathbb{C}$. For finding all possible invariants this has little consequence due to the following:

Fact 1. For polynomial maps ϕ and f as above, $f(\phi(x_1, \ldots, x_N)) = 0$ for all choices of $(x_1, \ldots, x_N)$ in an open subset of $\mathbb{R}^N$ if, and only if, $f(\phi(x_1, \ldots, x_N)) = 0$ for all choices of $(x_1, \ldots, x_N)$ in $\mathbb{C}^N$.

Thus while phylogenetic invariants express polynomial relationships that must hold for joint distributions arising from stochastically-meaningful parameter values, they are exactly the same relationships that hold for all complex parameter values.

Suppose we knew several polynomials in the set I, that evaluate to zero on $\phi(\mathbb{C}^N)$. From these, say $f_1(P), f_2(P), \ldots, f_k(P) \in I$, we can find many more, since for any choice of polynomials $h_i(P)$, the polynomial $\sum_{i=1}^{k} h_i(P) f_i(P)$ will then vanish wherever all the f_i do. Thus any such combination of invariants will also be an invariant. In the language of algebra, this means the set I of polynomials vanishing on $\phi(\mathbb{C}^M)$ forms an *ideal*. For a phylogenetic model, we call the collection of all invariants the *phylogenetic ideal*.

[4] A more optimistic hope would be that

$$P = (p_1, \ldots, p_M) \in \phi(\mathbb{C}^N) \text{ if, and only if, } f(p_1, \ldots, p_M) = 0 \text{ for all } f \in I,$$

but this is usually not possible. The common zero set of all $f \in I$ is closed, while $\phi(\mathbb{C}^M)$ may not be, and thus the zero set may contain additional points.

[5] Some writers refer to these merely as 'invariants,' reserving 'phylogenetic' for those invariants we refer to as *topologically informative*. We use 'phylogenetic invariant' to mean any invariant for a phylogenetic model.

But we must be more explicit about the role of the tree parameter, T, in a phylogenetic model. Even if we have fixed a model to consider, such as GM, the form of the parameterization map depends intimately on T. We signify this by denoting the parameterization map by ϕ_T, and its image by $\phi_T(\mathbb{C}^M)$. The phylogenetic ideal is the set of polynomials vanishing on this image, and so also depends on T. We typically denote the phylogenetic ideal by I_T, as we consider different trees. We omit from our notation a reference to the model, such as GM, since this is usually fixed throughout a discussion.

Since an ideal I is generally an infinite set of polynomials, to specify its elements we can ask for a list of *generators*, that is, a set of polynomials $\{f_1, f_2, \dots\}$ such that if $f \in I$ then $f = \sum_i h_i f_i$ for some choices of polynomials h_i. Fortunately, only finitely many generators are needed:

Fact 2. Any ideal of complex polynomials in M variables has a finite set of generators.

Thus, to find all invariants for a phylogenetic model and tree T, it is enough to determine a finite set of generators of the ideal I_T. For most ideals there is no canonical choice of a set of generators; different sets might generate the same ideal. In the phylogenetic setting we will of course prefer that our generators can be given a statistical explanation, such as the conditional independence interpretations of the edge and vertex invariants introduced earlier.

Given a collection S of polynomials in variables $P = (p_1, \dots, p_M)$, define the *algebraic variety* associated to S as

$$V(S) = \{P \in \mathbb{C}^M \mid f(P) = 0 \text{ for all } f \in S\}.$$

Thus the variety is simply the set of common zeros of the polynomials in S.

In particular, for phylogenetic models, we refer to $V_T = V(I_T)$, the common zero set of all phylogenetic invariants, as the *phylogenetic variety*. The phylogenetic variety will typically be larger than $\phi_T(\mathbb{C}^M)$, including points in the topological closure of the image of the parameterization. Thus the phylogenetic variety is made up of all 'joint-distributions' arising from complex parameter values, together with some additional points nearby.

When studying a model in the framework of algebraic geometry, finding generators for the phylogenetic ideal is certainly the most desirable goal. However, proving that one has found generators is often technically quite difficult, and a weaker result may be the best we can achieve.

Let V be an algebraic variety and I the ideal of all polynomials vanishing on V. Suppose S is some other set of polynomials having the same zero set as I, so that $V(S) = V$. Then we say S *defines* V *set-theoretically*. In such a circumstance $S \subset I$, but we may have $S \subsetneq I$, and even that S fails to generate I. While having a collection of set-theoretic defining polynomials for a variety does give us a way to test whether a point lies on a variety, we do not necessarily know all such tests unless we have generators of I.

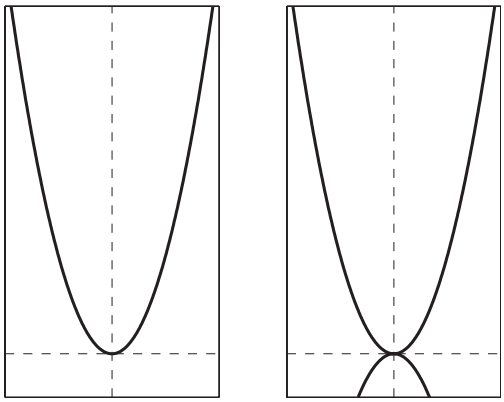

FIG. 4.8. The real points in varieties (a) defined by $p_1^2 - p_2 = 0$, or by $(p_1^2 - p_2)^2 = 0$, and (b) defined by $(p_1^2 - p_2)(p_1^2 + p_2) = 0$.

In order to clarify this terminology, we give a simple example, outside a phylogenetic setting. Consider the parameterization

$$\phi : \mathbb{C} \to \mathbb{C}^2$$
$$\phi(x) = (x, x^2).$$

The real points in the image of ϕ are shown in Fig. 4.8(a). It is easy to guess, correctly, that the ideal I of all polynomials vanishing on $\phi(\mathbb{C})$ is generated by the single polynomial $p_1^2 - p_2$. But notice that $(p_1^2 - p_2)^2$ also defines the variety set-theoretically, even though it does not generate the ideal. A third invariant is $p_1^4 - p_2^2 = (p_1^2 - p_2)(p_1^2 + p_2)$, which defines too large a variety, the union of the one of interest and its reflection below the p_1-axis. Although phylogenetic models with their many variables are necessarily more complicated, this simple example illustrates the main points: one might characterize ideal generators as the 'least complicated description' of a variety, and set-theoretic defining polynomials as a 'good description'. Other sets of polynomials have extraneous common zeros.

In principle, for a particular model on a particular tree, passing from a parameterization to an implicit description of a variety can be done by a computation involving variable elimination. This *implicitization problem* is described more fully in the excellent and accessible introduction to algebraic geometry [20], or more specifically in the phylogenetic setting by [37]. Computational algebra software implementing Gröbner basis algorithms, such as Maple, or the more specialized and powerful packages such as Macaulay2 [34] or Singular [35], can thus sometimes be used to explore invariants, form conjectures, and prove results.

However, several caveats on using computational algebra with phylogenetic models are in order. First, despite the impressive abilities of these packages,

the large number of variables involved in phylogenetic problems can make the computations intractable except for small trees and some of the less-complicated models. Second, the form of the invariants one finds this way can depend on computational choices that are made along the way, such as the term order necessary for any Gröbner basis computation. Therefore one will usually still want to find an interpretation, or natural construction, of the invariants produced computationally. Despite this, such computational explorations have played important roles in quite a few recent works focused on both finding and using invariants. Such software is an extremely valuable tool.

In many early papers on invariants, dimension counting was applied to determine how many invariants might be 'needed' for a particular model. If a model depended on N numerical parameters (with no redundancy), and gave a joint distribution with M entries, then the phylogenetic variety should be an N-dimensional object in M-dimensional space, i.e. of *codimension* $L = M - N$. Thus one might look for L phylogenetic invariants to define the variety set-theoretically.

Unfortunately, an algebraic variety of codimension L may require more than L set-theoretic defining polynomials. Although for some neighbourhood of any point there will be L polynomials defining the part of the variety in that neighbourhood, those polynomials may have additional common zeros outside of the neighbourhood that are not part of the variety. There may not be *any* set of L polynomials defining the variety globally.

This issue was first clearly brought up rather recently, in [37]. (See also the expository papers [38, 47].) In [66], as a consequence of the determination of all invariants for some group-based models, Sturmfels and Sullivant established that this issue did in fact arise for some standard phylogenetic models; previously given sets of invariants had many extraneous zeros. The authors argued strongly for the determining of the full ideal of invariants, or at least set-theoretic defining polynomials.

As a result of this history, one must be careful in interpreting literature that refers to 'complete sets of invariants which are algebraic generators' of the ideal. The concept of algebraic generators is a weaker one than set-theoretic defining polynomials, allowing extraneous zeros such as those in Fig. 4.8(b) when the variety of interest is (a). While such local defining polynomials might still be useful for future applications, it is likely that one needs some understanding of the locations of their extraneous zeros.

There are many open mathematical questions concerning phylogenetic ideals and varieties, some of which have been surveyed for algebraic geometers in [23]. Here we mention only one issue whose relevance will immediately be clear.

As mentioned, the vanishing of the invariants for a particular model and tree does not just distinguish joint distributions arising from parameter values that are probabilistically meaningful, but also those arising from complex parameters. This is not because of any lack of understanding of all invariants on our part, but rather due to the fundamental features of defining sets by the vanishing of polynomials. The field of *real algebraic geometry*, in which polynomial

inequalities as well as equalities play a role, would be a more appropriate setting in which to work if we hope to understand points coming from real parameter values. Although polynomial inequalities were used in both of the papers [13, 48] inaugurating the study of invariants, more recent works have not dealt with them. Real algebraic geometry presents greater technical difficulties than complex algebraic geometry, but it may provide greater understanding as well.

4.6 Invariants for specific models

Invariants have been found for phylogenetic models by many means, ranging from insightful observations, to exact algebraic computations, to more brute-force numerical computations.

Many papers focused on determining linear invariants for various models [12, 30, 31, 32, 42, 51, 65], partly because of the behaviour of linear invariants for rate-variation models that had been noted in [48] and will be discussed in Section 4.7. Other investigations, including [26, 27, 28, 29], found higher degree invariants. Already in [13] it was pointed out that some of these invariants encode statements of independence of substitutions in different parts of the tree, a theme that was further elaborated on in [21, 55].

Rather than survey these works in detail, we instead focus on some results obtained more recently. We hope this will provide a clearer overview of what invariants are and how they might be useful.

4.6.1 *Group-based models*

Group-based models, such as the Kimura 3-parameter model, and their submodels, such as the Jukes–Cantor and Kimura 2-parameter models, have a particularly nice mathematical structure which aids us in determining invariants. Since a full explanation could require a chapter in itself, we provide only an overview.

The key to analysing group-based models is the powerful tool of Fourier analysis. This was first recognized in Hendy's discovery of the Hadamard conjugation in [33, 41] for the 2-state symmetric model, where the underlying group is $\mathbb{Z}_2$. (See [40] for a more recent overview.) The relationship between the Kimura 3-parameter model and the group $\mathbb{Z}_2 \times \mathbb{Z}_2$, and the utilization of the associated Fourier transform, formed the basis of Evans and Speed's [24] insightful construction of invariants for the model. Fourier ideas were further explored for arbitrary group-based models in the work of Székely, Steel, and Erdős [67], which was then exploited for constructing invariants in [63]. See also [25]. Phylogenetic invariants for group-based models, then, appeared to be well-understood.

Recently, however, the question was considered of whether these constructions gave essentially all invariants: could one produce an explicit list of generators for the phylogenetic ideal for a group-based model? This was addressed by Sturmfels and Sullivant in [66].

The Fourier transform developed in the earlier works cited above amounts to a linear change of variables for the parameterization map, in both inputs

and outputs. The result of this transformation is that the complicated polynomial formulas for the parameterization map become quite simple: they can be given by monomials (one-term polynomials) in the transformed variables. Varieties parameterized by monomial functions are called *toric varieties* in algebraic geometry, and form a class that is particularly amenable to detailed analysis.

Using this, Sturmfels and Sullivant were able to show that all invariants for a particular tree could be constructed from invariants from the two smaller trees obtained by breaking an edge, together with some invariants associated to the edge itself. This 'breaking' or 'gluing' process reduced the problem of explicitly finding all invariants for an arbitrary tree to that for star trees, with only one internal node. Thus, after an analysis for the 3-leaf tree was completed, generators of the ideal for any binary tree could be explicitly given. We quote only a summary form of their result [66].

Theorem 4.1 *For a binary tree T, the ideal of phylogenetic invariants for the models $\mathcal{M}$ below is generated by the stochastic invariant, together with an explicit set of polynomials of the given degrees:*

$\mathcal{M} = 2$*-state symmetric, degree* 2*;*
$\mathcal{M} = 4$*-state Jukes–Cantor, degree* $1, 2, 3$*;*
$\mathcal{M} = 4$*-state Kimura* 2*-parameter, degree* $1, 2, 3, 4$*;*
$\mathcal{M} = 4$*-state Kimura* 3*-parameter, degree* $2, 3, 4$*.*

In addition to the explicit nature of the theorem, and the insight of the underlying analysis, there are two larger lessons to be drawn from these results.

First, the work shows that all invariants for group-based models arise from local features in the tree—from edges and nodes. As one considers trees with additional taxa, there will be larger sets of invariants, but their construction remains straightforward. Because the number of invariants needed to generate the phylogenetic ideal grows at least exponentially with the number of taxa, if invariants are to be useful for large trees, some local understanding of their meaning is valuable. Being able to tie generating invariants to specific topological features within a tree is likely to be essential for any application they may have.

Second, as mentioned in Section 4.5, it could be seen that for the 2-state symmetric model on a 4-leaf tree the 'complete sets of algebraic generators' of the invariants found in earlier works had many extraneous zeros. Indeed, the natural set of generators of the ideal of invariants for this model had more than the codimension number of polynomials in it, and any subset had extraneous zeros. This clearly showed that finding generators of the phylogenetic ideal, or at least set-theoretic defining polynomials, is necessary for adequate understanding of a phylogenetic variety.

Although we omit a detailed exposition of the precise form and construction of the invariants for group-based models, the 'Small Trees' web site [9] provides a valuable entryway for those interested in seeing or using them. It gives a compilation of invariants, Fourier transforms, and other information for trees of up to 5 taxa, with and without a molecular clock assumption. Input files for both Maple and Singular are helpfully provided.

4.6.2 *The general Markov model*

A separate thread of work on invariants was also undertaken recently, for the general Markov model, some of whose invariants were introduced in Sections 4.3 and 4.4. This model has many more parameters than the group-based models, and in studying it we lack the tool of Fourier analysis on a group. Nonetheless, fairly complete results have been obtained.

For the GM model, a single invariant for a 4-taxon tree was first given in [61]. The underlying idea was a suitable encoding of the 4-point condition for metric trees of [8], using log-det distances, building on an approach taken in [13]. Remarks in [59] point out that many additional invariants can be produced from the entries of certain matrix equations built from the joint distribution array. All these invariants depend only on two-dimensional marginalizations of the joint distribution (i.e. comparisons of sequences two at a time), as the underlying reasoning takes a generalized distance viewpoint.

The edge invariants for the GM model, which have been described in Section 4.3, are not inspired by any distance reasoning. Recall that they can be interpreted as statements of the independence of the substitution process on parts of the tree separated by an edge, conditioned on the state at some point along that edge.

For the 2-state GM model on a binary tree, the edge invariants in fact provide generators of the phylogenetic ideal, as was conjectured in [52] and proved in [6].

Theorem 4.2 *For the 2-state GM model on any n-leaf binary tree T, let P denote an n-dimensional array of indeterminants representing the joint distribution array. Then the ideal of phylogenetic invariants is generated by the stochastic invariant, together with all 3×3 minors of the matrix edge flattenings $\mathrm{Flat}_e(P)$ for all interior edges e of T.*

For instance in the 5-taxon tree example discussed at the end of Section 4.3, the 448 minors of size 3×3 of the two matrices shown, the edge invariants, provide a set of generators of the ideal.

Although the proof of this theorem requires mathematical techniques we will not discuss here, the result has a concrete, accessible interpretation: to each internal edge e of a tree we can associate both an explicit collection of cubic polynomials (the edge invariants for e) and a split of the taxa (into the two sets separated by e). These polynomials will be zero for any joint distribution arising from an n-taxon model on a tree inducing the same split of taxa. Moreover, these polynomials are essentially the only polynomial relationships that hold for all joint distributions arising from the fixed tree. Thus the structure of invariants for the GM model is determined by local features of the tree.

For the κ-state GM model, with $\kappa > 2$, our understanding is not quite as complete, but partial results again indicate a prominent role for invariants associated to local tree topology. The best current result is the following from [6].

Theorem 4.3 *Suppose a set of polynomials set-theoretically defining the variety associated to the GM model on a 3-taxon tree were given. Then an explicit*

construction will produce a set of polynomials set-theoretically defining the phylogenetic variety for any n-taxon binary tree.

Although this statement fails to highlight it, the construction of the explicit polynomials it refers to involves precisely the vertex invariants and edge invariants as discussed earlier. A large tree is viewed as many star trees joined together, and from invariants for each star tree, set-theoretic defining polynomials for the large tree are constructed.

We also note that while an understanding of set-theoretic defining polynomials for the 3-leaf tree is not complete, good partial results are available in [1] for the 4-state GM model.

Theorem 4.4 *Let S be the set of* 1728 *degree-*5 *polynomials, constructed as discussed in Section 4.4, which are invariants for the* 4*-state GM model on the* 3*-leaf tree. Then $V(S)$, the variety they define, is the union of the phylogenetic variety and possibly a set of extraneous zeros which lies in an explicitly describable set.*

The extraneous zeros mentioned in this last theorem can even be shown to be far from points on the phylogenetic variety arising from biologically-relevant parameter values.

We emphasize that the results for group-based and GM models parallel one another, in that all invariants ultimately arise from edges and nodes of the tree. Explicit polynomials tied to these features either generate the ideal or at least set-theoretically define the variety. However, the methods of proof are quite different. For group-based models, in addition to using the Fourier transform, the arguments are combinatorial in flavor and depend on an understanding of toric varieties. For the GM model, linear algebra and representations are the main ingredients.

4.6.3 *The strand symmetric model*

While the elegant mathematical structure of the group-based models facilitates an understanding of invariants, their restrictive assumptions are not always viewed as biologically realistic. While the GM model is also well-structured for understanding invariants, it might be considered to be too flexible, with too many parameters, for some phylogenetic applications. It is desirable, then, to look for biologically motivated models between these whose invariants can be successfully determined.

One potentially valuable one is the *strand symmetric* model introduced by Cassanellas and Sullivant in [10]. This 4-state model can be viewed as an amalgamation of a 2-state group-based model with a 2-state GM model, and thus its study can build on our understanding of each of those.

Specifically, with the fixed ordering of bases A,G,T,C, the model assumes a root distribution vector of the form

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_1, \pi_2),$$

so that frequency of any base matches its complement in the Watson–Crick pairing. This symmetry with respect to the pairing is also assumed for all Markov matrices on edges, so that they have the form

$$M_e = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ c & d & a & b \\ g & h & e & f \end{pmatrix}.$$

Since the rows of these matrices must sum to 1, there are 6 parameters introduced for each edge. Note that with this ordering of the bases the matrices have a block structure with 2×2 GM blocks arranged in a pattern reflecting the 2-state symmetric model.

As one might expect, the symmetry of this model leads to the existence of some linear invariants for any tree. Focusing next on the 3-taxon tree, a number of invariants of degree 3 and 4 can be constructed. However, it is not known whether these generate the phylogenetic ideal, or even set-theoretically define the phylogenetic variety, echoing the incompleteness of the corresponding result for the GM model. However, through the use of a computational algebra package, it can be seen that they generate all invariants of degree at most 4.

Finally, to handle trees relating more taxa, it is established that producing a set of invariants set-theoretically defining the variety for a 3-taxon tree would suffice to allow construction of invariants set-theoretically defining the variety for an arbitrary binary tree. This emphasizes once again that for those models for which we have made substantial progress in understanding invariants, we can tie particular invariants to particular local features of the tree.

4.6.4 *Stable base distribution models*

Another attempt to consider a biologically motivated model less general than the GM, but more general than group-based ones, appeared in [3]. The motivation was to understand what invariants might be valid for any model assuming a stable base distribution throughout the tree. In the course of this investigation, several nested models with this feature are formulated, including 1) an *algebraic time-reversible model* (ATR), which is similar to the GTR model but unlike the GTR has a polynomial parameterization map and, 2) a *stable base distribution model* (SBD) that assumes only that all Markov matrices fix the root distribution.

In the case of characters with 2 states, these models become the same, and generalize a model that had appeared earlier in [26]. In this simple situation the parameterization map is even explicitly invertible by a rational function; parameters can be recovered from a joint distribution by quite simple formulas. However, for a larger number of states our understanding is quite incomplete.

While some invariants are constructed for these models, little is understood about the full phylogenetic ideal or variety. Perhaps the most interesting result is a construction of a specific invariant that involves the hyperdeterminant of

a 3-dimensional array, making a connection between phylogenetic invariants and what mathematicians refer to as invariant theory. Though only of degree 6 for the 2-state model, unfortunately this invariant is of degree 408 in the 4-state case.

Part 2. Using Invariants

In this section we turn from questions of determining invariants for various models, to questions of how they might be used. Although invariants have had key roles in several contributions to theoretical understanding, for data analysis it is still less clear how they can be exploited. While their potential is attractive, much more needs to be done to develop ways to use them with data.

4.7 Invariants and statistical tests

In the decade following the first appearance of phylogenetic invariants in [13] and [48], many papers appeared building upon the idea. In particular, a number of these works dealt primarily with linear invariants for various models.

A compelling reason for the emphasis on linear invariants was the hope that they might be particularly useful for certain types of rate variation models. Suppose an invariant $f(P)$ for a specific model on a specific tree T is found which is linear and homogeneous (without constant term). Then since $f(c_1P_1 + \cdots + c_kP_k) = c_1f(P_1) + \cdots + c_kf(P_k)$, this polynomial will also vanish on any linear combination of joint distributions arising from the model. But linear combinations such as $c_1P_1 + \cdots + c_kP_k$ arise naturally when we consider *mixture models*, where sites are distributed among classes, and each class has its own set of parameters for the same model and tree. Then P_i represents the joint distribution for the ith class, and c_i the class size parameter. Thus a linear invariant for a model on a tree will also be a linear invariant for any *rates-across-sites* extension of that model on the same tree. We need not even make any assumptions about the nature of the distribution of sites among rate classes. This observation on linear invariants for mixture models holds for both discrete and continuous distributions of rates.[6]

If an invariant for a model is *topologically informative*, in the sense that it vanishes on all joint distributions arising from the model for some tree topologies and not others, then it could be the basis of a statistical test to distinguish between the topologies. Topologically-informative linear invariants, then, could give tests for topologies that would be insensitive to across-site rate variation. Tests of various sorts based on linear invariants were proposed in [11, 48], and investigated more thoroughly in [50].

Although higher degree invariants for a model are typically not invariants for rates-across-sites extensions, some attention was also given to how they might be used in a statistical framework. In [13], one of the quadratic invariants

[6] A higher degree invariant for a specific 2-class mixture model was first constructed in [29]. While this demonstrated that higher-degree invariants might be sought for mixture models, until recently it remained an isolated result.

constructed encoded an independence statement, that substitutions in one part of a tree were independent of those in another part of the tree separated from it by an edge. Thus the possibility of a statistical analysis based on 2-way contingency tables, as is typically done to test for independence, was suggested. This idea was pursued further in [55]. Using general formulas for multinomial distributions to estimate variances of quadratic invariants was suggested in [21]. However, as far as we know, no firmly-grounded statistical test based on general non-linear invariants has been suggested.

Several comparison studies [44, 45, 46] of the effectiveness of various phylogenetic inference methods included Lake's linear invariants. Using simulated data, Lake's method was found to be less efficient than many other methods, in that it required much longer data sequences to perform well. Note that Lake's method had been shown to be statistically consistent, so that provided data was in accord with the underlying model, as the length of data sequences approaches infinity the probability of inferring the correct tree approaches 1. Despite this theoretical strength, on sequences of a length typical for real data sets, Lake's invariants failed to reliably infer the correct tree even when no underlying model assumptions were violated.

In retrospect, this is not so surprising. Linear invariants only can test if a data point is in the smallest linear subspace containing the phylogenetic variety. Though higher degree invariants could potentially yield much more information than linear ones, a statistical framework for using them was largely lacking.

Indeed, how to use higher degree invariants in a statistically meaningful way is still an open question, and one needing exploration. There is evidence [9] that naive approaches to identifying topologies using all invariants can be effective on simulated data even with relatively short sequence length. Thus the inefficiency of Lake's linear invariants should not be interpreted as a sign that invariants in general are necessarily inefficient.

4.8 Invariants and maximum-likelihood

In current software, when phylogenetic inference is performed using a maximum-likelihood approach, the maximization of the likelihood function is undertaken by numerical search for optimal model parameters. For a possible tree topology, an attempt is made to find optimal numerical parameters such as base distributions, mutation rates, and edge lengths, and then the tree is varied and a new search for optimal parameters is undertaken.

Various algorithms can be used for the two aspects of this search (for numerical parameters and for topology), but rarely can one be certain that the true maximum has been located. For a fixed tree, a good algorithm will ensure locally optimal numerical parameters will be found, but the possibility of missing a global optimum remains. In addition, because the number of possible tree topologies will be quite large when the number of taxa is big, it may be impossible to consider all topologies, and so heuristic searches of tree spaces may overlook the optimal tree.

While many packages incorporate methods to avoid being trapped at non-global optima as they search, they generally come with no guarantee. Comparing the performance of one algorithm against another may shed some light on the issue, but cannot really give us full understanding if we have no way to verify that any maximum we have found is the true one.

Beginning with the work of Yang [69], a number of papers have sought to better understand the maximum-likelihood (ML) problem through exact optimization in simple settings.[7] In particular Chor and his collaborators introduced the use of phylogenetic invariants as an aid in this optimization problem.

To see why invariants might be useful for exact ML optimization, recall the construction of the likelihood function for a fixed n-leaf tree whose leaves are labeled by taxa. We first express the joint distribution of bases by an n-dimensional array P, as in Section 4.2. With variables $\mathbf{u} = (u_1, u_2, \ldots, u_L)$ representing the numerical parameters of the model, each entry of $P = P(\mathbf{u}) = (p_{i_1 i_2 \ldots i_n}(\mathbf{u}))$ is thus expressed by a polynomial parameterization function.

Given aligned sequences for the taxa, we record the observed distribution of bases as an n-dimensional array $\widehat{P} = (\widehat{p}_{i_1 i_2 \ldots i_n})$. The log-likelihood function is then

$$\ln L(\mathbf{u}) = \sum (\widehat{p}_{i_1 i_2 \ldots i_n}) \ln(p_{i_1 i_2 \ldots i_n}(\mathbf{u})).$$

To find maxima of this function, we can first look for critical points, where all partial derivatives are zero. Thus differentiating with respect to each variable u_j we obtain the system of equations

$$0 = \sum \frac{\widehat{p}_{i_1 i_2 \ldots i_n}}{p_{i_1 i_2 \ldots i_n}(\mathbf{u})} \frac{\partial p_{i_1 i_2 \ldots i_n}(\mathbf{u})}{\partial u_j}, \quad j = 1, \ldots, L.$$

Now since each $p_{i_1 i_2 \ldots i_n}(\mathbf{u})$ is a polynomial, these are rational equations. Clearing denominators, they give rise to a system of polynomial equations in the unknown parameters $\mathbf{u}$. If they can be solved, then among the solutions lie all local maxima of the likelihood function. Note that the polynomials $p_{i_1 i_2 \ldots i_n}(\mathbf{u})$ are typically of high degree (e.g. of degree approximately the number of edges in the tree), and clearing denominators could therefore lead to equations of very high degree.

While solving such a system of equations by hand is not usually possible, one might hope that a computer algebra package could handle it. Unfortunately, the polynomial system one obtains, even for a simple model on a small tree, may be intractable for current software.

However, this optimization problem can be reformulated as a constrained optimization problem that may be tractable. Rather than seek optimal parameters $\mathbf{u}$, we instead seek optimal values for the entries $p_{i_1 i_2 \ldots i_n}$ of P. We'd like to constrain P so that it lies in the image of the parameterization map, so we impose the slightly weaker condition that it lie in the phylogenetic variety. Thus we require that all phylogenetic invariants vanish on P. The ML problem

[7] Though this is often referred to as seeking *analytic* solutions to ML, we avoid that terminology as the methods are in fact generally algebraic.

becomes one of maximizing

$$\ln L(P) = \sum (\widehat{p}_{i_1 i_2 \ldots i_n}) \ln(p_{i_1 i_2 \ldots i_n})$$

subject to the constraints

$$f(P) = 0 \quad \text{for } f \in I_T.$$

Note that the model parameters do not appear here; we view the entries of P as the variables. Moreover, since the phylogenetic ideal I_T is finitely generated, only finitely many constraint equations $f_i(P) = 0$, $i = 1, \ldots, K$, are actually needed here.

Formulating this problem using Lagrange multipliers, all critical points are found by solving the system given by the K constraint equations together with the κ^n equations from the entries of

$$\nabla \ln L(P) = \sum_{i=1}^{K} \lambda_i \nabla f_i(P).$$

Explicitly, these last equations are simply

$$\frac{\widehat{p}_{i_1 i_2 \ldots i_n}}{p_{i_1 i_2 \ldots i_n}} = \sum_{i=1}^{K} \lambda_i \frac{\partial f_i}{\partial p_{i_1 i_2 \ldots i_n}}.$$

Though we again need to clear denominators to obtain polynomial equations, note that the resulting equations may well be of much lower degree then the ones obtained from the original parameter formulation of the ML problem, especially if the degrees of phylogenetic invariants are not that large.

This last observation gives some hope that with judicious use of a computer algebra system we might able to solve this constrained optimization problem. Indeed this is the case, at least for some small trees and simple models.

In [17] this approach was used to show that maximum likelihood estimation of trees could be quite ill-behaved. For a 2-state symmetric model on a 4-leaf tree, a number of examples of observed distributions $\widehat{P}$ were constructed for which the ML problem on a particular tree topology had a continuum of global maxima. For some of these, the global maxima even tied with a continua of global maxima for the other possible tree topologies as well. Proving these results for the specific examples required algebraic methods of solution of the above constrained optimization problem. The symmetry of the model results in some linear invariants which first allow a reduction in the number of variables $p_{i_1 i_2 \ldots i_n}$. Because the model is group-based, higher degree (quadratic) invariants could be constructed using the Fourier transform in the form of the Hadamard conjugation.

The paper [15] gives a more positive result on maximum likelihood, focusing on the 2-state symmetric model on a 3-taxon tree with a molecular clock, as had Yang in [69]. For this model, a linear invariant resulting from the molecular clock hypothesis is found through Hadamard conjugation. Using the constrained optimization formulation of the ML problem, the authors were able not only to

recover Yang's result on uniqueness of the ML optimum for this model on a fixed tree, but to extend it to allow variation in rates across sites, with mild restriction on the distribution of the rate parameter.

In [18, 19], the 2-state symmetric model with a molecular clock hypothesis is considered again, but now on 4-taxon trees. Hadamard conjugation again facilitates the derivation of invariants from the molecular clock hypothesis, though these must be derived separately for each of the possible rooted 4-taxon tree shapes, a 'fork' and a 'comb', and are quadratic rather than linear. The constrained optimization formulation of the ML problem is then solved, by a mix of insightful reductions and computer calculation. For the fork a unique maximum is found, whose coordinates can even be given as rational expressions in the entries of $\widehat{P}$. For the comb, the result is a bit more complicated, but the system is ultimately seen to have a finite number of solutions. However, all but one of these solutions is complex or outside the range $[0, 1]$, so again there is a unique maximum with statistical meaning.

In [16], this sort of analysis is pushed to a 4-state Jukes–Cantor model, on rooted 3-leaf trees. By working with transformed 'path-set' variables arising through Hadamard conjugation, rather than the variables $p_{i_1 i_2 \ldots i_n}$, the authors are able to avoid explicit use of constraint equations. Still, a symbolic algebra software package is needed to find critical points in the unconstrained formulation. They show that the ML problem has a finite number of optima, though some of the parameter values may not be meaningful in the context of the model.

Whenever a statistical model is parameterized through polynomial equations, one might take a similar algebraic approach to ML optimization. In [43], Hoşten, Khetan, and Sturmfels provide a general framework for using algebra to find exact solutions of ML problems. Computational approaches to both the constrained and unconstrained formulations are given. The authors further report that the constrained version generally performs better, though to take that approach requires one first finding model invariants, which of course may be quite difficult.

That paper also contains several phylogenetic calculations as examples. In one, for real data, the ML tree using a 4-state Jukes–Cantor model with 4 taxa is found, with the existence of a second local maximum established for that data also. This further indicates that multiple local maxima are a genuine issue in practical inference by maximum-likelihood. In another example, the result of [16] is reproved, this time in a constrained formulation.

The recent volume [53] provides a broader view of algebraic perspectives on statistics, with particular focus on applications to computational biology. Included in it is further background on the connections between algebra and general maximum-likelihood estimation.

4.9 Invariants and identifiability of complex models

While invariants were originally proposed for inferring trees from data, they can also be used to give theoretical results that such inference is possible. Separate

from the question of what inference method performs best for data analysis, is the more fundamental question concerning the limits of what can be inferred under perfect conditions.

A statistical model is said to be *identifiable* if from any joint distribution arising from the model it is possible to recover all parameters or, in other words, if the parameterization map of the model is injective. Identifiability is important because it plays a key role in proofs that methods of inference such as maximum likelihood are statistically consistent. If, for instance, two different tree topologies could give rise to the same joint distribution under some model, it is intuitively clear that inferring the 'correct' tree from data cannot be done reliably.

In practice, for phylogenetic models one must modify the strict notion of identifiability. For instance, allowing no substitutions to occur on an internal edge would lead to non-identifiability of the tree topology for 4-taxon trees, since each of the 3 fully-resolved 4-taxon trees as well as a 4-leaf star tree could all lead to the same joint distribution. Allowing too much substitution along internal edges, so that states become completely 'randomized' and uncorrelated in different parts of the tree, can also lead to loss of phylogenetic signal and non-identifiability of topology. Even when the tree parameter is identifiable for a model, numerical parameters may not be. For instance, for the GM model one can permute the states at an internal node of the tree, adjusting parameters appropriately, without changing the joint distribution [1, 14], so that numerical parameters are not identifiable unless one places additional restrictions on them. But while understanding the issues of non-identifiability mentioned so far is important, these are rather mild problems that can be dealt with by imposing biologically plausible assumptions on parameter values.

Identifiability of the tree parameter is often of primary interest in phylogenetics. For many basic models, such as the Jukes–Cantor, Kimura, or even GM, tree identifiability can be shown by first defining an appropriate phylogenetic distance, and then using the 4-point condition [8]. However, for models without a known distance formula, such as the covarion model [68], this approach is not possible. General mixture models, in which different classes of sites undergo substitutions according to different numerical parameter values for a model, but with the same tree parameter, also lack a distance. In both these situations tree identifiability has been an open question.

Note that while identifiability of the GTR+I+Γ model was shown in [54], the approach makes use of the assumption that the rate-parameters are described by a known distribution in such a way that the 4-point condition can still be applied. If the rate-parameter distribution is unknown for GTR+rates-across-sites model, then [64] established the topology is not identifiable for certain (non-explicit) parameter choices.

How general non-identifiability of a tree might be is quite important, both for knowing whether a particular model might be usable for inference, and for understanding under what circumstances tree inference might simply be impossible.

Phylogenetic invariants were recently used to study the problem of identifiability of the tree parameter for a variety of models in [4]. General theorems are produced that guarantee tree identifiability for most parameter choices for both the covarion model and many mixture models, provided the number of classes is small.

In order to study a variety of models at once, a substitution model is introduced that is much like the general Markov, but which allows λ states for the characters at internal nodes of the tree, and κ states at the leaves, with $\lambda \geq \kappa$. For DNA models with several classes, the states at the internal nodes might be indexed by pairs (i, j), where i refers to the base A,G,C,T and j to a rate-class, while at the leaves the states are simply the bases. Thus if there are n rate classes, then we have $\lambda = 4n$ states for all ancestral taxa, but only $\kappa = 4$ states for the currently extant taxa. The idea behind this is simply that while each site is in some rate-class, we cannot observe that class when data is collected; only the base can be recorded. The generality of this framework encompasses not only rates-across sites models, in which no site can change class, but also covarion models, where rate-class switching can occur.

While most invariants for such a model, even on a 4-leaf tree, are beyond our current knowledge, some can be found through a generalization of the edge invariant construction for the GM model. It can then be shown that these invariants are sufficient to identify the tree topology for *generic* choices of parameters, provided $\lambda < \kappa^2$. 'Generic' is given a precise meaning of 'all except those in a proper subvariety'. Since such a subvariety is necessarily of lower dimension than the parameter space, this means that if parameters are chosen randomly, according to any reasonable notion of randomness, they will be generic and the tree topology can be identified from the resulting joint distribution.

This result is for a model much more general than typically of interest in phylogenetics. Further arguments are given to show that when more usual mixture models are viewed as submodels of this general model they inherit identifiability of trees for generic parameters of their own.

In particular for κ-state models, even a GM+GM+$\cdots$+GM model, with a mixture of $\kappa - 1$ classes each described by the GM model but with unrelated numerical parameters, has identifiable tree topology for generic parameters. For DNA models, then, trees are identifiable for generic parameters of models with 3 unrelated GM classes. The result further specializes to a model such as the GTR, where a common rate matrix is assumed for the substitutions on all edges, allowing up to 3 classes of sites with scaled rates.

While the framework of invariants seems best suited to studying models with a finite number of rate-classes, much research literature refers to continuous distributions of rates. Indeed, the commonly-used GTR+Γ model assumes a continuous distribution. In fact, though, software implementations usually use discretized versions of Γ with only a few classes (although more than 3). Thus models with a finite number of rate-classes are common in practice.

It should be emphasized that there is no reason to believe identifiability for generic parameters should not hold for rate-class models with more than $\kappa - 1$

classes, provided the number of classes is not too large. The current restriction to $\kappa - 1$ classes is an artifact of having incomplete knowledge of all invariants for the models. A better understanding of what limits must be placed on the number of classes to preserve generic identifiability is still needed.

In addition to giving results on mixture models, [4] leads to establishing generic identifiability of the tree topology for certain covarion models, such as that of Tuffley and Steel [68] and extensions. Covarion models are biologically quite attractive in that they describe sites passing between being invariable and being free to vary as they evolve over a tree. However, identifiability of trees had not previously been established for them, despite their implementation in software [33].

For some of the results described here, such as for the covarion model and the GTR+rate-classes models, the underlying model is *not* one with a polynomial parameterization. These are inherently continuous time models, involving matrix exponentials in their parameterization formulas. Nonetheless, because they are submodels of a more general polynomially-parameterized model, they can be effectively studied through invariants.

Another investigation [5] of invariants for mixture models has focused on the GM+I model, with 2 classes, one evolving according to GM and the other held invariable. Although identifiability of the tree for generic parameters in this model follows from [4], a focus on this more specific model allows additional invariants to be found, giving a refined analysis. Note that some questions of identifiability for this model had been studied previously in [7], in which it was shown the tree was not identifiable from marginalizations of the joint distribution to 2 taxa (i.e. from pairwise sequence comparisons).

An interesting consequence of studying invariants for GM+I is a set of explicit formulas that can recover the proportion of invariable sites with any given base from the joint distribution. For the more restrictive Kimura 3-parameter model with invariable sites, such a formula was found in [62] by a rather different argument using 'capture/recapture' reasoning. For the GM+I model an understanding of the invariants naturally leads to determinantal formulas to recover these parameters. For example, in the 2-state case on a 4 taxon tree, with states 0 and 1, the proportion of invariable characters of state 0 is given as a quotient:

$$\pi_0^I = \frac{\begin{vmatrix} p_{0000} & p_{0001} & p_{0010} \\ p_{0100} & p_{0101} & p_{0110} \\ p_{1000} & p_{1001} & p_{1010} \end{vmatrix}}{\begin{vmatrix} p_{0101} & p_{0110} \\ p_{1001} & p_{1010} \end{vmatrix}}.$$

Here subscripts indicating states corresponding to the taxa ordered as a, b, c, d, where the tree has split $ab|cd$. Similar formulas are valid for the 4-state characters, or even κ-state.

Note that such formulas are far from unique, since they can be modified by the addition of any invariant for the model without affecting the value the

formula will yield when evaluated at a distribution. Nonetheless, there is a possibility that such formulas might be useful for quick estimation of parameters from data.

Identifiability by means of invariants also appeared in [2], which focused on the use of invariants only for quartets (subsets of 4 taxa) to determine a fit of n data sequences to a tree. Although the precise results require some technical conditions, they can be roughly summarized as indicating that while quartet invariants can indicate a unique n-taxon tree, additional invariants are needed to assure the n-dimensional joint distribution is fit well by the model. This clarifies the loss of information inherent in quartet methods of inference.

4.10 Other directions

4.10.1 *A tree construction algorithm*

A first step toward a novel invariant-based inference method was taken by Eriksson in [22], with a software implementation for DNA sequence data. The underlying idea uses only the edge invariants for the GM model. Following an algorithmic approach reminiscent of neighbour joining, the method iteratively builds a tree by finding good taxa, or clades, to join together, and thus has good running times.

In the initial step, all splits that separate two taxa from the rest are considered. If all the edge invariants for a hypothetical split come close to vanishing, then that is evidence that the two taxa should be joined. However, evaluating these invariants would simply be a test that the corresponding flattening of the observed distribution is close to a rank 4 matrix. Thus, rather than actually evaluate the many edge invariants for each flattening, the algorithm instead uses a numerical approach to determine how close each flattening is to a matrix with rank 4.

This problem of measuring how well a matrix can be approximated by one of fixed rank is well understood, provided closeness is measured by the Frobenius (i.e. L^2 on matrix entries) norm. The singular value decomposition of matrices provides a good numerical approach both to finding such approximations, and measuring error. Thus the algorithm avoids both the issues of how to use the large number of invariants associated to one edge to get a combined measure of support for that edge, and how one would interpret such a measure in a statistically meaningful way.

Although the performance of Eriksson's SVD method on simulated data was not as good as neighbour joining or maximum-likelihood, as a first attempt it gave several reasons to be hopeful. First, the simulation studies were in some sense biased against the new method: data was simulated according to a more restricted model than the GM model underlying the SVD algorithm, so that one might expect the generality of the GM model allowed too much flexibility in parameters for optimal tree recovery. It would be interesting to see how the algorithm's performance compares on simulated data that violates some of the common assumptions of the competing methods. For data arising without stable

base frequencies throughout evolution, or with substitution rates on different edges of the tree varying substantially, the GM model may be valid where something like the GTR is not. Indeed, in such a situation the SVD method could be proved to be statistically consistent, unlike standard implementations of other methods, which do not allow such flexibility in models.

Second, the SVD method is based only on consideration of edge invariants, and not of vertex invariants. In a sense, it is dealing with a model even more general than GM, by placing no assumptions at all on how substitutions occur around the time of speciation events. While one might expect better performance if vertex invariants are somehow utilized, it is unfortunately not immediately clear how to do so. There is no simple analogue of the SVD for determining best approximations of 3-dimensional tensors of specified rank, so new ideas are likely to be needed.

Although more needs to be done to develop this approach, there is also much potential to do so. The focus on the relationship of invariants to local tree structure, as well as the introduction of the SVD to provide an alternative to naive evaluation for 'near-vanishing' of polynomials, can guide future work.

4.10.2 *Invariants for gene order models*

In [56, 57, 58], a new direction in the application of invariants was given by Sankoff and Blanchette, to inferring phylogenetic trees from gene order data. Not only are parsimony approaches to inference in this setting computationally slow even for quite small trees, but they can also produce incorrect results if there are large differences in branch lengths in the tree. Since invariants are based on a model, and are designed to 'ignore' specific parameter values such as branch lengths, they might provide a useful new approach.

First a simplified probabilistic model is given to describe gene order data with n genes. Focusing on any particular gene, the various states for the model are the possible genes that might be its successor in the ordering. Assuming equal probabilities of all such changes on a given edge of the tree, an $(n-1)$-state model generalizing the Jukes–Cantor one is produced. Thus linear invariants are well understood, and can be explicitly produced for a small tree and small n. From simulation using parameters inferred from the data, distributions for the values of these invariants can be produced, and significance levels assigned to the values they produce when evaluated on the data. For real data, the method produces plausible results, in line with a parsimony approach focusing only on adjacent genes. While there is possibly some improvement in inference, examples are too few to be conclusive.

As the authors noted, little had been done with probabilistic models for evolving gene order, and the simple model they used is only a very rough approximation that might be improved. They also investigated only linear invariants, the construction of which was already well known for this model, noting their insensitivity to rate variation. Producing a more sophisticated model and determining its invariants might well enable better inference, though how difficult that might be is unclear.

4.11 Concluding remarks

Much progress has been made in understanding invariants of various phylogenetic models. Only recently has it been possible to claim we know all invariants for some models, or even a large number of invariants for models general enough to include those commonly used in inference. For group-based models and the GM model our knowledge is now extensive, and a pleasing and potentially useful relationship between invariants and local tree features has emerged. Even for certain very general mixture models, we have learned of some non-linear invariants that are topologically informative.

Moreover invariants have proved their usefulness in addressing two fundamental theoretical issues. They played an important role in investigating the possibility of multiple maxima of the likelihood functions, making it possible to formulate the problem as one of constrained optimization so that exact solutions could be found. They also were the key tool in establishing the identifiability of trees for general mixture models, with a small number of classes, for generic parameters.

How invariants might be useful in practical inference is now a question ready for renewed exploration. Earlier disappointments in the performance of linear invariants should not be discouraging, since that small subclass of invariants offers little insight into how higher degree ones might perform. For naive approaches to using invariants for inference to be developed into useful and well-founded methods, we need to find both good ways of evaluating a large number of invariants, and good statistical approaches to judging whether the results are near to zero. But as the SVD algorithm has shown, we might let invariants guide our thinking yet use other computational ideas in developing an inference method.

Simply put, we do not yet know how to use invariants to address practical problems. Although their potential seems clear, the development of ways to use invariants, either heuristically or in well-founded statistical tests, needs the attention of a wider group of researchers.

References

[1] Allman, E. S., and Rhodes, J. A. (2003). Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences*, **186**, 113–144.

[2] Allman, E. S. and Rhodes, J. A. (2004). Quartets and parameter recovery for the general Markov model of sequence mutation. *Applied Mathematics Research eXpress*, **2004**(4), 107–131.

[3] Allman, E. S. and Rhodes, J. A. (2006). Phylogenetic invariants for stationary base composition. *Journal of Symbolic Computation*, **41**(2), 138–150.

[4] Allman, E. S., and Rhodes, J. A. (2006). The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *Journal of Computational Biology*, **13**(5), 1101–1113. `arXiv:q-bio.PE/0511009`.

[5] Allman, E. S. and Rhodes, J. A. (2007). Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites. `arXiv:q-bio:PE/0702050`.

[6] Allman, E. S. and Rhodes, J. A. (2007). Phylogenetic ideals and varieties for the general Markov model. To appear in, *Advances in Applied Mathematics*, `arXiv:math.AG/0410604`.

[7] Baake, E. (1998). What can and what cannot be inferred from pairwise sequence comparisons? *Mathematical Biosciences*, **154**(1), 1–21.

[8] Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In *Mathematics in the Archeological and Historical Sciences*, pp. 387–395. Edinburgh University Press, Edinburgh.

[9] Casanellas, M., Garcia, L. D., and Sullivant, S. (2005). Catalog of small trees. In *Algebraic Statistics for Computational Biology* (ed. L. Pachter and B. Sturmfels), pp. 291–304. Cambridge University Press, Cambridge. `http://www.math.tamu.edu/~lgp/small-trees/`.

[10] Casanellas, M. and Sullivant, S. (2005). The strand symmetric model. In *Algebraic Statistics for Computational Biology* (ed. L. Pachter and B. Sturmfels), pp. 305–321. Cambridge University Press, Cambridge.

[11] Cavender, J. A. (1989). Mechanized derivation of linear invariants. *Molecular Biology and Evolution*, **6**, 301–316.

[12] Cavender, J. A. (1991). Necessary conditions for the method of inferring phylogeny by linear invariants. *Mathematical Biosciences*, **103**, 69–75.

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

[14] Chang, J. T. (1996). Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical Biosciences*, **137**(1), 51–73.

[15] Chor, B., Hendy, M., and Penny, D. (2001). 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 Computer Science*, pp. 204–213. Springer, Berlin.

[16] Chor, B., Hendy, M., and Snir, S. (2006). Maximum likelihood Jukes-Cantor triplets: Analytic solutions. *Molecular Biology and Evolution*, **23**(3), 626–632. `arXiv:q-bio.PE/0505054`.

[17] Chor, B., Hendy, M. D., Holland, B. R., and Penny, D. (2000). Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Molecular Biology and Evolution*, **17**, 1529–1541.

[18] Chor, B., Khetan, A., and Snir, S. (2003). Maximum likelihood on four taxa phylogenetic trees: Analytic solutions. *RECOMB'03*, pp. 76–83. ACM Press, New York.

[19] Chor, B. and Snir, S. (2004). Molecular clock fork phylogenies: Closed form analytic maximum likelihood solutions. *Systematic Biology*, **53**(6), 963–967.

[20] Cox, D., Little, J., and O'Shea, D. (1997). *Ideals, Varieties, and Algorithms* (2nd edn.). Springer-Verlag, New York.

[21] Drolet, S. and Sankoff, D. (1990). Quadratic tree invariants for multivalued characters. *Journal of Theoretical Biology*, **144**, 117–129.

[22] Eriksson, N. (2005). Tree construction using singular value decomposition. In *Algebraic Statistics for Computational Biology* (ed. L. Pachter and B. Sturmfels), pp. 347–358. Cambridge University Press, Cambridge.

[23] Eriksson, N., Ranestad, K., Sturmfels, B., and Sullivant, S. (2004). Phylogenetic algebraic geometry. In *Projective Varieties with Unexpected Properties; Siena, Italy*, (Eds. Ciro Ciliberto, Antony V. Geramita, Brian Harbourne, Rosa Maria Miró–Roig, and Kristian Ranestrad) pp. 237–256. de Gruyter, Berlin. `arXiv:math.AG/0407033`.

[24] Evans, S. N. and Speed, T. P. (1993). Invariants of some probability models used in phylogenetic inference. *Annals of Statistics*, **21**(1), 355–377.

[25] Evans, S. N. and Zhou, X. (1998). Constructing and counting phylogenetic invariants. *Journal of Computational Biology*, **5**(4), 713–724.

[26] Ferretti, V., Lang, B. F., and Sankoff, D. (1994). Skewed base compositions, asymmetric transition matrices, and phylogenetic invariants. *Journal of Computational Biology*, **1**(1), 77–92.

[27] Ferretti, V. and Sankoff, D. (1993). The empirical discovery of phylogenetic invariants. *Advances in Applied Probability*, **25**(2), 290–302.

[28] Ferretti, V. and Sankoff, D. (1995). Phylogenetic invariants for more general evolutionary models. *Journal of Theoretical Biology*, **173**, 147–162.

[29] Ferretti, V. and Sankoff, D. (1996). A remarkable nonlinear invariant for evolution with heterogeneous rates. *Mathematical Biosciences*, **134**(1), 71–83.

[30] Fu, Y. (1995). Linear invariants under Jukes' and Cantor's one-parameter model. *Journal of Theoretical Biology*, **173**, 339–352.

[31] Fu, Y. and Li, W. (1992). Construction of linear invariants in phylogenetic inference. *Mathematical Biosciences*, **109**, 201–228.

[32] Fu, Y. and Li, W. (1992). Necessary and sufficient conditions for the existence of linear invariants in phylogenetic inference. *Mathematical Biosciences*, **108**, 203–218.

[33] Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution*, **18**(5), 866–873.

[34] Grayson, D. R. and Stillman, M. E. (2002). Macaulay2, a software system for research in algebraic geometry. Available at `http://www.math.uiuc.edu/Macaulay2/`.

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

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

[37] Hagedorn, T. R. (2000). Determining the number and structure of phylogenetic invariants. *Advances in Applied Mathematics*, **24**(1), 1–21.

[38] Hagedorn, T. R. and Landweber, L. F. (2000). Phylogenetic invariants and geometry. *Journal of Theoretical Biology*, **205**, 365–376.

[39] Hendy, M. D. (1989). The relationship between simple evolutionary tree models and observable sequence data. *Systematic Zoology*, **38**, 310–321.

[40] Hendy, M. D. (2005). Hadamard conjugation: An analytic tool for phylogenetics. In *Mathematics of Evolution and Phylogeny* (ed. O. Gascuel), pp. 143–177. Oxford University Press, Oxford.

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

[42] Hendy, M. D. and Penny, D. (1996). Complete families of linear invariants for some stochastic models of sequence evolution, with and without the molecular clock assumption. *Journal of Computational Biology*, **3**(1), 19–31.

[43] Hoşten, S., Khetan, A., and Sturmfels, B. (2005). Solving the Likelihood Equations. *Foundations of Computational Mathematics. The Journal of the Society for the Foundations of Computational Mathematics*. **5**(4), 389–407. `arXiv:math.ST/0408270`.

[44] Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Systematic Biology*, **44**(1), 17–48.

[45] Huelsenbeck, J. P. and Hillis, D. M. (1993). Success of phylogenetic methods in the four-taxon case. *Systematic Biology*, **42**(3), 247–264.

[46] Jin, L. and Nei, M. (1990). Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular Biology and Evolution*, **7**(1), 82–102.

[47] Kim, J. (2000). Slicing hyperdimensional oranges: The geometry of phylogenetic estimation. *Molecular Phylogenetics and Evolution*, **17**(1), 58–75.

[48] Lake, J. A. (1987). A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Molecular Biology and Evolution*, **4**(2), 167–191.

[49] Landsberg, J. M. and Manivel, L. (2004). On the ideals of secant varieties of Segre varieties. *Foundations of Computational Mathematics*, **4**(4), 397–422.

[50] Navidi, W. C., Churchill, G. A., and von Haeseler, A. (1993). Phylogenetic inference: Linear invariants and maximum likelihood. *Biometrics*, **49**(2), 543–555.

[51] Nguyen, T. and Speed, T. P. (1992). A derivation of all linear invariants for a nonbalanced transversion model. *Journal of Molecular Evolution*, **35**, 60–76.

[52] Pachter, L. and Sturmfels, B. (2004). Tropical geometry of statistical models. *Proceedings of the National Academy of Sciences, USA*, **101**(46), 16132–16137 (electronic).

[53] Pachter, L. and Sturmfels, B. (ed.) (2005). *Algebraic Statistics for Computational Biology*. Cambridge University Press, Cambridge.

[54] Rogers, J. S. (2001). Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. *Systematic Biology*, **50**(5), 713–722.

[55] Sankoff, D. (1990). Designer invariants for large phylogenies. *Molecular Biology and Evolution*, **7**(3), 255–269.

[56] Sankoff, D. and Blanchette, M. (1999). Phylogenetic invariants for genome rearrangements. *Journal of Computational Biology*, **6**(3/4), 431–445.

[57] Sankoff, D. and Blanchette, M. (1999). Probability models for genome rearrangements and linear invariants for phylogenetic inference. In *Proceeedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99)*, pp. 302–309. ACM Press, New York.

[58] Sankoff, D. and Blanchette, M. (2000). Comparative genomics via phylogenetic invariants for Jukes-Cantor semigroups. In *Stochastic models (Ottawa, ON, 1998)*, pp. 399–418. American Mathematical Society, Providence.

[59] Semple, C. and Steel, M. (1999). Tree representations of non-symmetric group-valued proximities. *Advances in Applied Mathematics*, **23**(3), 300–321.

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

[61] Steel, M. (1994). Recovering a tree from the leaf colourations it generates under a Markov model. *Applied Mathematics Letters*, **7**(2), 19–23.

[62] Steel, M., Huson, D., and Lockhart, P. J. (2000). Invariable sites models and their uses in phylogeny reconstruction. *Systematic Biology*, **49**(2), 225–232.

[63] 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. *New Zealand Journal of Botany*, **31**(31), 289–296.

[64] Steel, M., Székely, L. and Hendy, M. D. (1994). Reconstructing trees from sequences whose sites evolve at variable rates. *Journal of Computational Biology*, **1**(2), 153–163.

[65] Steel, M. A. and Fu, Y. X. (1995). Classifying and counting linear phylogenetic invariants for the Jukes-Cantor model. *Journal of Computational Biology*, **2**(1), 39–47.

[66] Sturmfels, B. and Sullivant, S. (2005). Toric ideals of phylogenetic invariants. *Journal of Computational Biology*, **12**(2), 204–228. `arXiv:q-bio.PE/0402015`.

[67] Székely, L. A., Steel, M. A., and Erdős, P. L. (1993). Fourier calculus on evolutionary trees. *Advances in Applied Mathematics*, **14**(2), 200–210.

[68] Tuffley, C. and Steel, M. (1998). Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences*, **147**(1), 63–91.

[69] Yang, Z. (2000). Complexity of the simplest phylogenetic estimation problem. *Proceedings of the Royal Society of London B: Biological Sciences*, **267**, 109–116.