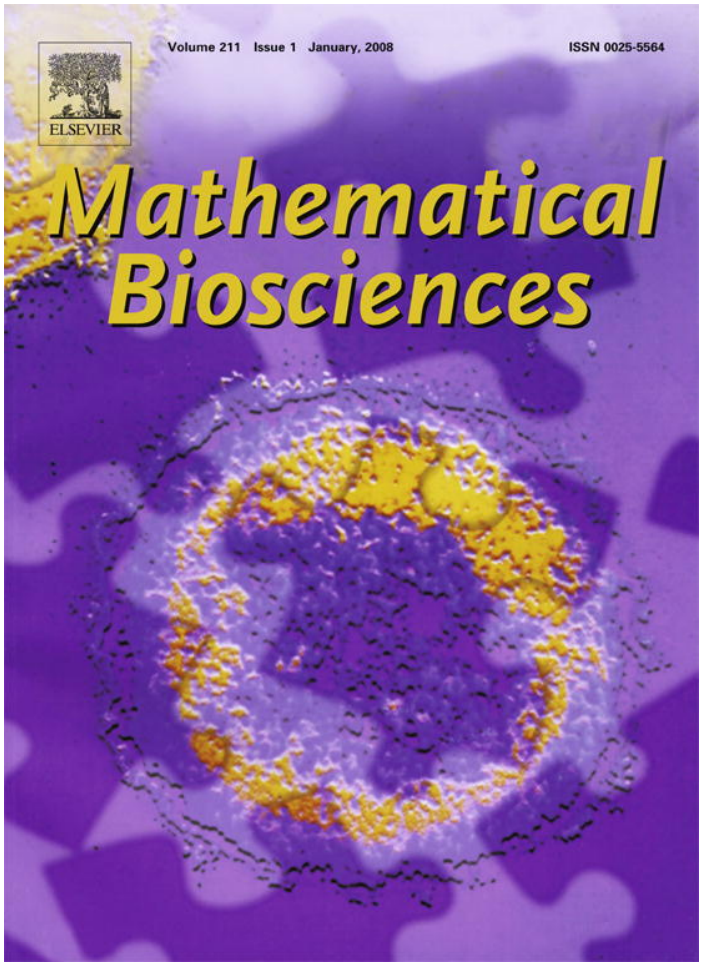
Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article was published in an Elsevier journal. The attached copy is furnished to the author for non-commercial research and education use, including for instruction at the author's institution, sharing with colleagues and providing to institution administration.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

Available online at www.sciencedirect.com
ScienceDirect
Mathematical Biosciences 211 (2008) 18–33
Mathematical Biosciences
www.elsevier.com/locate/mbs

Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites ☆

Elizabeth S. Allman, John A. Rhodes *

Department of Mathematics and Statistics, University of Alaska Fairbanks, P.O. Box 756660, Fairbanks, AK 99775, USA

Received 24 February 2007; received in revised form 14 July 2007; accepted 13 September 2007
Available online 25 September 2007

Abstract

The general Markov plus invariable sites (GM + I) model of biological sequence evolution is a two-class model in which an unknown proportion of sites are not allowed to change, while the remainder undergo substitutions according to a Markov process on a tree. For statistical use it is important to know if the model is identifiable; can both the tree topology and the numerical parameters be determined from a joint distribution describing sequences only at the leaves of the tree? We establish that for generic parameters both the tree and all numerical parameter values can be recovered, up to clearly understood issues of 'label swapping'. The method of analysis is algebraic, using phylogenetic invariants to study the variety defined by the model. Simple rational formulas, expressed in terms of determinantal ratios, are found for recovering numerical parameters describing the invariable sites.
© 2007 Elsevier Inc. All rights reserved.

Keywords: Phylogenetics; Invariable site model; Identifiability; Phylogenetic invariants

☆ This research was supported in part by the Institute for Mathematics and its Applications, with funds provided by the National Science Foundation. We thank the IMA for its hospitality.
* Corresponding author.
E-mail addresses: e.allman@uaf.edu (E.S. Allman), j.rhodes@uaf.edu (J.A. Rhodes).

0025-5564/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.mbs.2007.09.001

Author's personal copy

1. Introduction

If a model of biological sequence evolution is to be used for phylogenetic inference, it is essential that the model parameters of interest – certainly the tree parameter and usually the numerical parameters – be identifiable from the joint distribution of states at the leaves of the tree. Though often unstated, the assumption that model parameters are identifiable underlies the use of both Maximum Likelihood and Bayesian inference methods; identifiability of parameters is prerequisite to establishing statistical consistency of these methods. As increasingly complicated models, incorporating across-site rate variation, covarion structure, or other types of mixtures, are implemented in software packages, there is a real possibility that non-identifiability could confound data analysis. Unfortunately, our theoretical understanding of this issue lags well behind current phylogenetic practice.

One natural approach to proving the identifiability of the tree topology relies on the definition of a phylogenetic distance for the model, and the 4-point condition of Buneman [1]. For instance, Steel [2] used the log-det distance to establish the identifiability of the tree topology under the general Markov model and its submodels. Such a distance-based argument shows additionally that 2-marginalizations of the full joint distribution suffice to recover the tree parameter, since distances require only two-sequence comparisons. Once the tree has been identified, the numerical parameters giving rise to a joint distribution for the general Markov model can be determined by an argument of Chang [3].

While distance measures have been developed for GTR models with rate variation [4,5], these require that one know the rate distribution completely, and identifiability of the rate distribution itself has yet to be addressed. Identifiability of the popular GTR + I + Γ model of sequence evolution was considered in [6], but there are significant gaps in the argument of that paper, as was pointed out to us by Ané [7]. Thus, for general mixture models, and even rates-across-sites models where no appropriate definition of a distance is known, proving the identifiability of the tree parameter requires a different approach.

The alternative viewpoint of algebraic geometry is used in [8], to show the generic identifiability of the tree parameter for the covarion model of [9] and for certain mixture models with a small number of classes. Though this result is far more general than previous identifiability results, it still fails to cover the type of rate-variation models currently in common use for data analysis, and does not address identifiability of numerical parameters at all. Much more study of the identifiability question is needed.

In this paper, we focus on the *general Markov plus invariable sites*, GM + I, model of sequence evolution, a model that encompasses the GTR + I model that is of more immediate interest to practitioners. Note that previous work on GM + I by Baake [10] focused on non-identifiability. In that paper parameter choices for the 2-state GM + I model on two distinct 4-taxon trees are constructed that give rise to the same pairwise joint distributions (2-marginals). As both sets of parameters have 50% invariable sites, this shows that the identifiability of the tree parameter cannot generally hold on the basis of 2-sequence comparisons, even if the distribution of rate factors is known. Furthermore, it implies that a well-behaved phylogenetic distance (computed from 2-marginals) cannot be defined for this model, as existence of such a distance would imply tree identifiability.

Author's personal copy

Here we prove that all parameters for the GM + I model are indeed identifiable, through 4-sequence comparisons. By identifiable, we mean *generically identifiable* in a geometric sense: for a fixed tree, the set of numerical parameters for which the joint distribution could have arisen from either (a) a different tree, or (b) a 'significantly different' (in a sense to be made clear later) choice of numerical parameters on the same tree, is of strictly lower dimension than that of the full numerical parameter space. (For a concrete example of generic identifiability, recall the results of Steel and Chang on the general Markov model: assumptions that the Markov edge matrices M_e have determinant $\neq 0, 1$ and that the distribution of states at the root has strictly positive entries ensure identifiability of all parameters. These are generic conditions.) Thus for natural probability distributions on the parameter space, with probability one a choice of parameters is generic.

Although identifiability of the tree parameter for GM + I follows from more general results in [8], that paper did not consider identifiability of numerical parameters. Our arguments here are tailored to GM + I and yield stronger results addressing numerical parameters as well as the tree. Our approach is again based on the determination of *phylogenetic invariants* for the model. While the invariants described in [8] are invariants for more general models than GM + I, the ones given in this paper apply only to GM + I and its submodels, and are of much lower degree.

As a byproduct of the development of these GM + I invariants, we are led to rational formulas for recovering all the parameters related to the invariable sites (proportion of invariable sites and stationary distribution of these invariable sites) from a joint distribution. Indeed, these formulas are crucial to our identification of numerical parameters.

These formulas can be viewed as GM + I analogs of the formulas for the proportion of invariable sites in group-based + I models that were found by the capture–recapture argument of [11]. In the group-based setting, those formulas were developed into a heuristic means of estimating the proportion of invariable sites from data without performing a full tree inference. This has been implemented in `SplitsTree4` [12]. At this time, it remains unclear whether a useful heuristic can be found for the formulas presented in this paper.

Recently, Jayaswal et al. [13] have described a software implementation for maximum likelihood estimation under the GM + I model under consideration here. We refer to that paper for remarks on the features of the model that make it an attractive one for certain biological data sets.

Since our algebraic methods at times employ computational commutative algebra software packages, and these tool are not commonly used in the phylogenetics literature, we have included sample code in Appendix A.

2. The GM + I model

Let T denote an *n-taxon tree*, by which we mean a tree with n leaves labeled by the taxa $a_1, a_2, \ldots, a_n$ and all internal vertices of valence at least 3. We say T is *binary* if all internal nodes have valence exactly 3.

We begin by describing the parameterization of the κ-state GM + I model of sequence evolution along T, where $\kappa = 4$ corresponds to usual models of DNA evolution. The *class size parameter* δ denotes the probability that any particular site in a sequence is invariable: conceptually, the flip of a biased coin weighted by δ determines if a site is allowed to undergo state transitions. If a

Author's personal copy

site is invariable, it is assigned state $i \in [\kappa] = \{1, 2, \ldots, \kappa\}$ with probability $\pi_{\mathrm{I}}(i)$. Here $\boldsymbol{\pi}_{\mathrm{I}} = (\pi_{\mathrm{I}}(1), \ldots, \pi_{\mathrm{I}}(\kappa))$ is a vector of non-negative numbers summing to 1 giving the state distribution for invariable sites.

All sites that are not invariable mutate according to a common set of parameters for the GM model, though independently of one another. For these sites, we associate to each node (including leaves) of T a random variable with state space $[\kappa]$. Choosing any node r of T to serve as a root, and directing all edges away from r, let T_r denote the resulting directed tree T. A *root distribution* vector $\boldsymbol{\pi}_{\mathrm{GM}} = (\pi_{\mathrm{GM}}(1), \ldots, \pi_{\mathrm{GM}}(\kappa))$, with non-negative entries summing to 1, has entries $\boldsymbol{\pi}_{\mathrm{GM}}(j)$ specifying the probability that the root variable is in state j. For each directed edge $e = (v \to w)$ of T_r, let M_e be a $\kappa \times \kappa$ Markov matrix, so that $M_e(i, j)$ specifies the conditional probability that the variable at w is in state j given that the variable at v is in state i. Thus entries of all M_e are non-negative, with rows summing to 1.

For the GM + I model on an n-taxon tree T with edge set E, the stochastic parameter space $S \subset [0, 1]^N$ is of dimension $N = 1 + (\kappa - 1) + (\kappa - 1) + \mid E \mid \kappa(\kappa - 1) = 2\kappa - 1 + \mid E \mid \kappa(\kappa - 1)$. The parameterization map giving the joint distribution of the variables at the leaves of T is denoted by

$$\begin{aligned} \phi_T : S &\to [0, 1]^{\kappa^n}, \\ \mathbf{s} &\mapsto P. \end{aligned}$$

We view P as an n-dimensional $\kappa \times \ldots \times \kappa$ array, with dimensions corresponding to the ordered taxa $a_1, a_2, \ldots, a_n$, and with entries indexed by the states at the leaves of T. The entries of P are polynomial functions in the parameters $\mathbf{s}$ explicitly given by

$$P(i_1, \ldots, i_n) = \delta\epsilon(i_1, i_2, \ldots i_n)\boldsymbol{\pi}_{\mathrm{I}}(i_1) + (1 - \delta) \sum_{(j_v) \in \mathcal{H}} \left(\boldsymbol{\pi}_{\mathrm{GM}}(j_r) \prod_e M_e(j_{v_i}, j_{v_f}) \right). \tag{1}$$

Here $\epsilon(i_1, i_2, \ldots i_n)$ is 1 if all i_j are equal and 0 otherwise, the product is taken over all edges $e = (v_i \to v_f) \in E$, and the sum is taken over the set of all possible assignments of states to nodes of T extending the assignment $(i_1, \ldots, i_n)$ to the leaves: if V is the set of vertices of T then

$$\mathcal{H} = \{(j_v) \in [\kappa]^{|V|} | j_v = i_k \text{ if } v \text{ is a leaf labeled by } a_k\}.$$

For notational ease, the entries of P, the *pattern frequencies*, are also denoted by $p_{i_1 \ldots i_n} = P(i_1, \ldots, i_n)$.

We note that while a root r was chosen for the tree in order to explicitly describe the GM portion of the parameterization of our model, the particular choice of r is not important. Under mild additional restrictions on model parameters, changing the root location corresponds to a simple invertible change of variables in the parameterization. (See [14,15], or [16] for details.) This justifies our slight abuse of language in referring to the GM or GM + I model on T, rather than on T_r, and we omit future references to root location.

Note that Eq. (1) allows us to more succinctly describe any $P \in \mathrm{Im}(\phi_T)$ as

$$P = (1 - \delta)P_{\mathrm{GM}} + \delta P_{\mathrm{I}}, \tag{2}$$

where P_{GM} is an array in the image of the GM parameterization map on T and $P_{\mathrm{I}} = \mathrm{diag}(\boldsymbol{\pi}_{\mathrm{I}})$ is an n-dimensional array whose off-diagonal entries are zeros and whose diagonal entries are those of π_{I}.

Author's personal copy

3. Model identifiability

We now make precise the various concepts of identifiability of a phylogenetic model. To adapt standard statistical language to the phylogenetic setting, for a fixed set A of n taxa and $\kappa \geqslant 2$, consider a collection $\mathcal{M}$ of pairs (T, ϕ_T), where T is an n-taxon tree with leaf labels A, and $\phi_T : S_T \to [0,1]^{\kappa^n}$ is a parameterization map of the joint distribution of pattern frequencies for the model on T. We say *the tree parameter is identifiable* for $\mathcal{M}$ if for every $P \in \cup_{(T,\phi_T)\in\mathcal{M}} \mathrm{Im}(\phi_T)$, there is a unique T such that $P \in \mathrm{Im}(\phi_T)$. We say that *numerical parameters are identifiable on a tree T* if the map ϕ_T is injective, that is if for every $P \in \mathrm{Im}(\phi_T)$ there is a unique $\mathbf{s} \in S_T$ with $\phi_T(\mathbf{s}) = P$. We say the *model $\mathcal{M}$ is identifiable* if the tree parameter is identifiable, and for each tree the numerical parameters are identifiable.

It is well known that such a definition of identifiability is too stringent for phylogenetics. First, unless one restricts parameter spaces, there is little hope that the tree parameter be identifiable: one need only think of any standard model on a binary 4-taxon tree in which the Markov matrix parameter on the internal edge is the identity matrix. Any joint distribution arising from such a parameter choice could have as well arisen from any other 4-taxon tree topology.

Even if such 'special' parameter choices are excluded so the tree parameter becomes identifiable, identifiability of numerical parameters also poses problems, as noted by Chang [3]. For example, consider the 3-taxon tree with the GM model. Then multiple parameter choices give rise to the same joint distribution since the labeling of the states at the internal node can be permuted in $\kappa!$ ways, as long as the Markov matrix parameters are adjusted accordingly [15]. The occurrence of this sort of 'label-swapping' non-identifiability in statistical models with hidden (unobserved) variables is well known, but is not of great concern. However, even for this model more subtle forms of non-identifiability can occur, in which infinitely many parameter choices lead to the same joint distribution. These arise from singularities in the model, and can be avoided by again restricting parameter space. Such 'generic' conditions for the GM model have already been mentioned in the introduction.

We therefore refine our notions of identifiability. Because we are concerned primarily with model where the maps ϕ_T are given by polynomials, we give a formulation appropriate to that setting. Recall that given any collection $\mathcal{F}$ of polynomials in N variables, their common zero set,

$$V(\mathcal{F}) = \{z \in \mathbb{C}^N | f(z) = 0 \text{ for all } f \in \mathcal{F}\},$$

is the *algebraic variety* defined by $\mathcal{F}$. If the algebraic variety is a proper subset of $\mathbb{C}^N$, then it is said to be *proper*.

Definition 1. Let $\mathcal{M}$ be a model on a collection of n-taxon trees, as defined above.

(1) *We say* the tree parameter is generically identifiable *for $\mathcal{M}$ if for each tree T there exists a proper algebraic variety X_T with the property that*
$P \in \cup_{(T,\phi_T)\in\mathcal{M}} \phi_T(S_T \setminus X_T)$ *implies* $P \in \phi_T(S_T \setminus X_T)$ *for a unique T.*

(2) *We say that* numerical parameters are generically locally identifiable on a tree *T if there is a proper algebraic variety Y_T such that for all $\mathbf{s} \in S_T \setminus Y_T$, there is a neighborhood of $\mathbf{s}$ on which ϕ_T is injective.*

Author's personal copy

(3) *We say the* model $\mathcal{M}$ is generically locally identifiable *if the tree parameter is generically identifiable, and for each tree the numerical parameters are generically locally identifiable.*

Note that the notion of 'generic' here is used to mean 'for all parameters but those lying on a proper subvariety of the parameter space,' and such a variety is necessarily of lower dimension than the full parameter space. Using the standard measure on the parameter space, viewed as a subset of $\mathbb{R}^N$, this notion thus also implies 'for all parameters except those in a set of measure 0'.

In the important special case of parameterization maps defined by polynomial formulas, such as that for the GM + I model, generic local identifiability of numerical parameters is equivalent to the notion in algebraic geometry of the map ϕ_T being *generically finite*. In this case, there exists a proper variety Y_T and an integer k, the degree of the map ϕ_T, such that restricted to $S_T \setminus Y_T$ the map ϕ_T is not only locally injective but also k-to-1: that is, if $\mathbf{s} \in S_T \setminus Y_T$ and $P = \phi_T(\mathbf{s})$, then the fiber $\phi_T^{-1}(P)$ has cardinality k.

Because of the label swapping issue at internal nodes, for the GM model and GM + I on an n-taxon tree T with vertex set V, fibers of generic points will always have cardinality at least $\kappa!(|V| - n)$. Thus for these models, the best we can hope for is generic local identifiability of the model (both tree and numerical parameters) where the generic fiber has exactly this cardinality. That in fact is what we establish in the next section.

4. Generic identifiability for the GM + I model

We begin our arguments by determining some phylogenetic invariants for the GM + I model. The notion of a phylogenetic invariant was introduced by Cavender and Felsenstein [17] and Lake [18], in the hope that phylogenetic invariants might be useful for practical tree inference. Their role here, in proving identifiability, is more theoretical but illustrates their value in analyzing models.

For a parameterization ϕ_T given by polynomial formulas on domain $S_T \subseteq \mathbb{R}^N$, we may uniquely extend to a polynomial map with domain $\mathbb{C}^N$, given by the same polynomial formulas, which we again denote by $\phi_T : \mathbb{C}^N \to \mathbb{C}^{\kappa^n}$.

Remark 2. Extending parameters to include complex values is solely for mathematical convenience, as algebraic geometry provides the natural setting for our viewpoint. The collection of stochastic joint distributions (arising from the original stochastic parameter space) is a proper subset of $\mathrm{Im}(\phi_T)$.

The *phylogenetic variety*, V_T, is the smallest algebraic variety in $\mathbb{C}^{\kappa^n}$ containing $\phi_T(\mathbb{C}^N)$, *i.e.*, the closure of the image of ϕ_T under the Zariski topology,

$$V_T = \overline{\mathrm{Im}(\phi_T)} \subseteq \mathbb{C}^{\kappa^n}.$$

Remark 3. V_T coincides with the closure of $\mathrm{Im}(\phi_T) = \phi_T(\mathbb{C}^N)$ under the usual topology on $\mathbb{C}^{\kappa^N}$. However, while $V_T \cap [0,1]^{\kappa^n}$ contains the closure of $\phi_T(S_T)$ under the usual topology, these need not be equal.

Author's personal copy

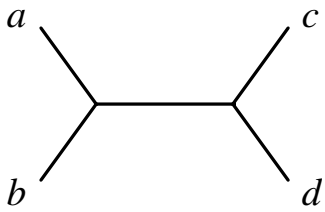


Fig. 1. The 4-taxon tree $T_{ab|cd}$.

Let $\mathbb{C}[P]$ denote the ring of polynomials in the κ^n indeterminates $\{p_{i_1\ldots i_n}\}$. Then the collection of all polynomials in $\mathbb{C}[P]$ vanishing on V_T forms a prime ideal I_T. We refer to I_T as a *phylogenetic ideal*, and its elements as *phylogenetic invariants*. More explicitly, a polynomial $f \in \mathbb{C}[P]$ is a phylogenetic invariant if, and only if, $f(P_0) = 0$ for every $P_0 \in \phi_T(\mathbb{C}^{\kappa^n})$, or equivalently, if, and only if, $f(P_0) = 0$ for every $P_0 \in \phi_T(S_T)$.

As we proceed, we consider first the special case of 4-taxon trees. We highlight the $\kappa = 2$ case, in part to illustrate the arguments for general κ more clearly, and in part because we can go further in understanding the 2-state model.

Consider the 4-taxon binary tree $T_{ab|cd}$, with taxa a, b, c, d as shown in Fig. 1.

Suppose that P is a $2 \times 2 \times 2 \times 2$ pattern frequency array, whose indices correspond to states $[2] = \{1, 2\}$ at the taxa in alphabetical order. Then the internal edge e of T defines the split $ab \mid cd$ in the tree, and we define the *edge flattening* F_e of P at e, a $2^2 \times 2^2$ matrix, by

$$F_e = \begin{pmatrix} p_{1111} & p_{1112} & p_{1121} & p_{1122} \\ p_{1211} & p_{1212} & p_{1221} & p_{1222} \\ p_{2111} & p_{2112} & p_{2121} & p_{2122} \\ p_{2211} & p_{2212} & p_{2221} & p_{2222} \end{pmatrix}. \tag{3}$$

Notice that the rows of F_e are indexed by the states at $\{ab\}$ and the columns by states at $\{cd\}$. The flattening F_e is intuitively motivated by considering a 'collapsed' model induced by e: taxa a and b are grouped together forming a single variable $\{ab\}$ with 4 states, and the grouping $\{cd\}$ forms a second variable with 4 states.

This construction can be generalized in a natural way: suppose T is an n-taxon tree, and P a $\kappa \times \cdots \times \kappa$ array with indices corresponding to the taxa labeling the leaves of T. Then for any edge e in T, we can form from P the matrix F_e of size $\kappa^{n_1} \times \kappa^{n_2}$, where n_1 and n_2 are the cardinalities of the two sets of taxa in the split induced by e.

From [16] (for a more expository presentation, see also [19]), we have:

Theorem 4. *For the 2-state GM model on a binary n-taxon tree T, the phylogenetic ideal I_T is generated by all 3×3 minors of all edge flattenings F_e of P. Moreover, for the κ-state GM model on an n-taxon tree T, the phylogenetic ideal I_T contains all $(\kappa+1) \times (\kappa+1)$ minors of all edge flattenings of P.*

Using this result, we can deduce some elements of the phylogenetic ideal for the GM + I model for any number of taxa $n \geqslant 4$ and any number of states $\kappa \geqslant 2$.

Proposition 5 (Phylogenetic invariants for GM + I).

Author's personal copy

(1) *For the* 4*-taxon tree* $T_{ab|cd}$ *and the* 2*-state GM + I model, the cubic determinantal polynomials*

$$f_1 = \begin{vmatrix} p_{1112} & p_{1121} & p_{1122} \\ p_{1212} & p_{1221} & p_{1222} \\ p_{2112} & p_{2121} & p_{2122} \end{vmatrix} \quad \text{and} \quad f_2 = \begin{vmatrix} p_{1211} & p_{1212} & p_{1221} \\ p_{2111} & p_{2112} & p_{2121} \\ p_{2211} & p_{2212} & p_{2221} \end{vmatrix}$$

are phylogenetic invariants. These are the two 3×3 *minors of the matrix flattening* $F_{ab|cd}$ *of Eq. (3) that do not involve either of the entries* p_{1111} *or* p_{2222}.

(2) *More generally, for* $n \geqslant 4$ *and* $\kappa \geqslant 2$, *consider the* κ*-state GM + I model on an n-taxon tree T. Then for each edge e of T, all* $(\kappa+1) \times (\kappa+1)$ *minors of the flattening* F_e *of P that avoid all entries* $p_{ii\ldots i}, i \in [\kappa]$ *are phylogenetic invariants.*

Proof. We prove the first statement in detail. From Eq. (2), for any $P = \phi_T(s)$ we have $P = (1-\delta)P_{\mathrm{GM}} + \delta P_{\mathrm{I}}$, where P_{GM} is a 4-dimensional table arising from the GM model on T and $P_{\mathrm{I}} = \mathrm{diag}(\pi_{\mathrm{I}})$ is a diagonal table with entries giving the distribution of states for the invariable sites. Flattening these tables with respect to the internal edge of the tree, we obtain

$$F_{ab|cd} = (1-\delta)F_{\mathrm{GM}} + \delta F_I = (1-\delta)\begin{pmatrix} \tilde{p}_{1111} & \tilde{p}_{1112} & \tilde{p}_{1121} & \tilde{p}_{1122} \\ \tilde{p}_{1211} & \tilde{p}_{1212} & \tilde{p}_{1221} & \tilde{p}_{1222} \\ \tilde{p}_{2111} & \tilde{p}_{2112} & \tilde{p}_{2121} & \tilde{p}_{2122} \\ \tilde{p}_{2211} & \tilde{p}_{2212} & \tilde{p}_{2221} & \tilde{p}_{2222} \end{pmatrix} + \delta \begin{pmatrix} \pi_I(1) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \pi_I(2) \end{pmatrix}. \tag{4}$$

By Theorem 4, all 3×3 minors of F_{GM} vanish. Since the 'upper right' and 'lower left' minors of $F_{ab|cd}$ are the same as those of F_{GM}, up to a factor of $(1-\delta)^3$, they also vanish.

Straightforward modifications to this argument give the general case. □

For arbitrary n, κ, the GM + I model should have many other invariants than those found here. Among these is, of course, the stochastic invariant

$$f_s(P) = 1 - \sum_{\mathbf{i} \in [\kappa]^n} p_{\mathbf{i}}.$$

In the simplest interesting case of the GM + I model, however, we have the following computational result.

Proposition 6. *The phylogenetic ideal for the* 2*-state GM + I model on the* 4*-taxon tree* $T_{ab|cd}$ *of Fig. 1 is generated by* f_s *and the minors* f_1, f_2 *above*;

$$I_T = \langle f_s, f_1, f_2 \rangle.$$

Proof. A computation of the Jacobian of the parameterization $\phi_T : S \subset \mathbb{C}^{13} \to \mathbb{C}^{24}$ shows it has full rank at some points, and so V_T is of dimension 13. If $I = \langle f_s, f_1, f_2 \rangle$, then $I \subseteq I_T$. Another computation shows that I is prime and of dimension 13. Thus, necessarily $I = I_T$. (The code for these computations is given in Appendix A.) □

Author's personal copy

Let $V_{ab|cd}, V_{ac|bc}, V_{ad|bc}$ be the varieties for the 2-state GM + I models for the three 4-taxon binary tree topologies, with corresponding phylogenetic ideals $I_{ab|cd}, I_{ac|bd}, I_{ad|bc}$. Of course Proposition 6 gives generators for each of these ideals – two 3×3 minors of the flattenings of P appropriate to those tree topologies, along with f_s. A computation (see Appendix A) shows that these three ideals are distinct. Therefore the three varieties are distinct, and their pairwise intersections are proper subvarieties. Thus for any parameters $\mathbf{s}$ not lying in the inverse image of these subvarieties, T is uniquely determined from $\phi_T(\mathbf{s})$. Thus we obtain

Corollary 7. *For the 2-state GM + I model on binary 4-taxon trees, the tree parameter is generically identifiable.*

As $\dim(V_{ab|cd}) = 13$, and the parameter space for ϕ_T is 13 dimensional, we also immediately obtain that the map ϕ_T is generically finite. This yields

Corollary 8. *For the 2-state GM + I model on a binary 4-taxon tree, numerical parameters are generically locally identifiable.*

Note that this approach does not yield the cardinality of the generic fiber of the parameterization map, which is also of interest. We will return to this issue in Theorem 13.

Further computations show that $\dim(V_{ab|cd} \cap V_{ac|bd} \cap V_{ad|bc}) = 11$. As this intersection contains all points arising from the GM + I model on the 4-taxon star tree, which is an 11-parameter model, this is not surprising. In fact, one can verify computationally that the ideal $I_{ab|cd} + I_{ac|bd} + I_{ad|bc}$ is the defining prime ideal of the star-tree variety. We also note that the ideal $I_{ab|cd} + I_{ac|bd}$ decomposes into two primes, both of dimension 11. Thus the variety defined by this ideal has two components, one of which is the variety for the star tree.

In principle, the ideal I_T of all invariants for the GM + I model on an arbitrary tree T can be computed from the parameterization map ϕ_T via an elimination of variables using Gröbner bases [20]. However, if all invariants for the κ-state GM model on T are known, they can provide an alternate approach to finding I_T which, while still proceeding by elimination, should be less computationally demanding.

To present this most simply, we note that because our varieties lie in the hyperplane described by the stochastic invariant, it is natural to consider their projectivizations, lying in $\mathbb{P}^{\kappa^n - 1}$ rather than $\mathbb{C}^{\kappa^n}$. The corresponding phylogenetic ideals, which we denote by J_T, are generated by the homogeneous polynomials in I_T, and do not contain the stochastic invariant. Conversely, I_T is generated by the elements of J_T together with the stochastic invariant.

In addition, we need not restrict ourselves to the GM model, but rather deal with any phylogenetic model parameterized by polynomials.

Proposition 9. *Suppose $\tilde{\phi}_T : \mathbb{C}^N \to \mathbb{C}^{\kappa^n}$ is a parameterization map for some phylogenetic model $\mathcal{M}$ on T, with corresponding homogeneous phylogenetic ideal $\widetilde{J}_T$. Let*

$$\phi_T : \mathbb{C}^N \times \mathbb{C}^\kappa \to \mathbb{C}^{\kappa^n}$$

be the parametrization map for the $\mathcal{M} + I$ model given by

$$\phi_T(\mathbf{s}, (\delta, \boldsymbol{\pi}_I)) = (1 - \delta)\tilde{\phi}_T(\mathbf{s}) + \delta \mathrm{diag}(\boldsymbol{\pi}_I).$$

Author's personal copy

Let P' denote the collection of all indeterminate entries of P except those in $P_{eq} = \{p_{ii\ldots i} \mid i \in [\kappa]\}$. Then the homogeneous phylogenetic ideal J_T for the $\mathcal{M}+I$ model on T is $J_T = (\widetilde{J}_T \cap \mathbb{C}[P'])\mathbb{C}[P]$. Thus J_T can be computed from $\widetilde{J}_T$ by elimination of the variables in P_{eq}.

Proof. Extend the parameterization maps $\tilde{\phi}_T, \phi_T$ to parameterizations of cones by introducing an additional parameter,

$$\widetilde{\Phi}_T(\mathbf{s}, t) = t\tilde{\phi}_T(\mathbf{s}),$$
$$\Phi_T(\mathbf{s}, (\delta, \boldsymbol{\pi}_{\mathrm{I}}), t) = t\phi_T(\mathbf{s}, (\delta, \boldsymbol{\pi}_{\mathrm{I}})).$$

Then $\mathrm{Im}(\Phi_T) = \mathbb{C}^{\kappa} \times \mathrm{proj}(\mathrm{Im}(\widetilde{\Phi}_T))$, where $\mathbb{C}^{\kappa}$ corresponds to coordinates in P_{eq} and 'proj' denotes the projection map from P-coordinates to P'-coordinates. As J_T is the ideal of polynomials vanishing on $\mathrm{Im}(\Phi_T)$, and $\widetilde{J}_T \cap \mathbb{C}[P']$ the ideal vanishing on $\mathrm{proj}(\mathrm{Im}(\widetilde{\Phi}_T))$, the result follows. □

Using this, in Appendix A we give an alternate computation to show both part (1) of Proposition 5, and Proposition 6. While this computation is quite fast, a more naive attempt to find GM + I invariants directly from the full parameterization map using elimination was unsuccessful, demonstrating the utility of the proposition. Moreover, we can use this proposition to compute all 2-state GM + I invariants on the 5-taxon binary tree as well. This leads us to

Conjecture 10. *On an n-taxon binary tree, the ideal of homogeneous invariants for the 2-state GM + I model is generated by those 3×3 minors of edge flattenings that do not involve the variables $p_{11\ldots1}$ and $p_{22\ldots2}$, together with the stochastic invariant.*

Although we are unable to determine all GM + I invariants for the 4-taxon tree for general κ, using only those described in Proposition 5 we can still obtain identifiability results through a modified argument.

Proposition 11. *For the κ-state GM + I model on binary 4-taxon trees, $\kappa \geqslant 2$, the tree parameter is generically identifiable.*

Proof. By the argument leading to Corollary 7, it is enough to show the varieties $V_{ab|cd}, V_{ac|bd}$, and $V_{ad|bc}$ are distinct. Considering, for example, the first two, we can show that the varieties $V_{ab|cd}$ and $V_{ac|bd}$ are distinct, by giving an invariant $f \in I_{ac|bd}$ and a point $P_0 \in V_{ab|cd}$ such that $f(P_0) \neq 0$.

Using Proposition 5, we pick an invariant $f \in I_{ac|bd}$ as follows: in the flattening $F_{ac|bd}$ according to the split $ac \mid bd$, choose any collection of $\kappa+1$ ac-indices with distinct a and c states, *e.g.*, $\{12, 13, \ldots, 1\kappa, 21, 23\}$. Using the same set as bd-indices, this determines a $(\kappa+1) \times (\kappa+1)$-minor f.

We pick $P_0 = \phi_{T_{ab|cd}}(\mathbf{s})$ using the parameterization of Eq. (1) by making a specific choice of parameters $\mathbf{s}$. On $T_{ab|cd}$, with the root r located at one of the internal nodes, choose parameters $\mathbf{s}$ as follows: let $\boldsymbol{\pi}_{\mathrm{GM}}, \boldsymbol{\pi}_{\mathrm{I}}$ be arbitrary but with all entries of $\boldsymbol{\pi}_{\mathrm{GM}}$ positive. Pick any $\delta \in [0, 1)$. For the four terminal edges choose M_e to be the $\kappa \times \kappa$ identity matrix I_{κ}. For the single internal edge e of T, choose any Markov matrix M_e with all positive entries. For such parameters, the entries of the joint distribution $P_0 = \phi_{T_{ab|cd}}(\mathbf{s})$ are zero except for the pattern frequencies p_{iijj}, where the states at the leaves a and b agree and the states at the leaves c and d agree. Since the entries of M_e and the root distributions are positive, each of the $p_{iijj} > 0$.

Author's personal copy

But considering the flattening $F_{ac|bd}$ of $P_0 = \phi_{T_{ab|cd}}(\mathbf{s})$ with respect to the 'wrong' topology $T_{ac|bd}$, we observe that the κ^2 non-zero entries p_{iijj} of $F_{ac|bd}$ all lie on the diagonal of $F_{ac|bd}$, in the positions with ij as both ac-index and bd-index. Furthermore, by our choice of f, a subset of them forms the diagonal of the submatrix whose determinant is f. Therefore $f(P_0) \neq 0$. □

Proposition 12 (Recovery of invariable site parameters).

(1) *For the* 4*-taxon tree* $T_{ab|cd}$ *and the* 2*-state GM + I model, suppose* $P = \phi_T(\mathbf{s})$. *Then generically the parameters in* $\mathbf{s}$ *related to invariable sites can be recovered from* P *by the following formulas*:

$$\delta = \frac{|A_1| + |A_2|}{|B|}, \quad \boldsymbol{\pi}_{\mathrm{I}} = \frac{1}{|A_1| + |A_2|}(|A_1|, |A_2|),$$

where $B = \begin{pmatrix} p_{1212} & p_{1221} \\ p_{2112} & p_{2121} \end{pmatrix}$,

$$A_1 = \begin{pmatrix} p_{1111} & p_{1112} & p_{1121} \\ p_{1211} & p_{1212} & p_{1221} \\ p_{2111} & p_{2112} & p_{2121} \end{pmatrix}, \quad A_2 = \begin{pmatrix} p_{1212} & p_{1221} & p_{1222} \\ p_{2112} & p_{2121} & p_{2122} \\ p_{2212} & p_{2221} & p_{2222} \end{pmatrix}.$$

(2) *More generally, for the* κ*-state GM + I model on* $T_{ab|cd}$, *the invariable site parameters can be recovered from a generic point in the image of the parameterization map by rational formulas of the form*

$$\delta = \frac{\sum_{i\in[\kappa]}|A_i|}{|B|}, \quad \boldsymbol{\pi}_{\mathrm{I}} = \frac{1}{\sum_{i\in[\kappa]}|A_i|}(|A_1|, |A_2|, \ldots, |A_n|).$$

Here $|\, B \,|$ *is any* $\kappa \times \kappa$ *minor of* $F_{ab|cd}$ *that omits the all rows and columns indexed by ii, and* $|\, A_i \,|$ *is the* $(\kappa+1)\times(\kappa+1)$ *minor obtained by including all rows and columns chosen for* B *and in addition the ii row and ii column.*

Proof. We give the complete argument in the case $\kappa = 2$ first. For a joint distribution $P \in \mathrm{Im}(\phi_T)$, write $F_{ab|cd} = (1-\delta)F_{\mathrm{GM}} + \delta F_{\mathrm{I}}$ as in Eq. (4). Since A_1 is the 'upper left' 3×3 submatrix of $F_{ab|cd}$, using linearity properties of the determinant, and that all 3×3 minors of F_{GM} evaluate to zero, we observe that

$$\begin{aligned} |\, A_1 \,| &= (1-\delta)^3 \begin{vmatrix} \tilde{p}_{1111} & \tilde{p}_{1112} & \tilde{p}_{1121} \\ \tilde{p}_{1211} & \tilde{p}_{1212} & \tilde{p}_{1221} \\ \tilde{p}_{2111} & \tilde{p}_{2112} & \tilde{p}_{2121} \end{vmatrix} + \begin{vmatrix} \delta\pi_{\mathrm{I}}(1) & 0 & 0 \\ 0 & (1-\delta)\tilde{p}_{1212} & (1-\delta)\tilde{p}_{1221} \\ 0 & (1-\delta)\tilde{p}_{2112} & (1-\delta)\tilde{p}_{2121} \end{vmatrix} \\ &= \delta\pi_{\mathrm{I}}(1) \begin{vmatrix} (1-\delta)\tilde{p}_{1212} & (1-\delta)\tilde{p}_{1221} \\ (1-\delta)\tilde{p}_{2112} & (1-\delta)\tilde{p}_{2121} \end{vmatrix}. \end{aligned}$$

Thus we have $|\, A_1 \,| = \delta\pi_{\mathrm{I}}(1)\,|\, B \,|$. Now, if $|\, B \,| \neq 0$, then

Author's personal copy

$$\delta\pi_{\mathrm{I}}(1) = \frac{|A_1|}{|B|}.$$

As $|B|$ does not vanish on all of V_T, we have a rational formula to compute $\delta\pi_{\mathrm{I}}(1)$ for generic points on V_T.

Similarly, since A_2 is the 'lower right' submatrix of $F_{ab|cd}$, then

$$\delta\pi_{\mathrm{I}}(2) = \frac{|A_2|}{|B|}.$$

Adding these together, we obtain the stated rational expression for δ.

Assuming additionally the generic condition that $\delta \neq 0$, then we find

$$\boldsymbol{\pi}_{\mathrm{I}} = \left(\frac{|A_1|}{|A_1| + |A_2|}, \frac{|A_2|}{|A_1| + |A_2|}\right).$$

Thus the parameters $\delta, \boldsymbol{\pi}_{\mathrm{I}}$ are generically identifiable for GM + I on T.

One readily sees the argument above can be modified for arbitrary κ. □

Note that when $\kappa > 2$ the above proposition gives many alternative rational formulas for the invariable site parameters, as there are many options for choosing the matrix B.

We now obtain our main result.

Theorem 13. *The κ-state GM + I model on n-taxon binary trees, with $n \geqslant 4, \kappa \geqslant 2$, is generically locally identifiable. Furthermore, for an n-taxon tree with V vertices, the fibers of generic points of V_T under the parametrization map have cardinality $\kappa!(|V| - n)$. Thus for generic points, label swapping at internal nodes is the only source of non-identifiability.*

Proof. Suppose T is an n-taxon tree with $P = \phi_T(\mathbf{s})$. Choose some subset of 4 taxa, say $\{a, b, c, d\}$, and suppose the induced quartet tree is $T_{ab|cd}$. Then P_{abcd}, the 4-marginalization of P, is easily seen to be of the form $P_{abcd} = \phi_{T_{ab|cd}}(\mathbf{s}_{abcd})$ *where* $\mathbf{s}_{abcd} = g(\mathbf{s})$ and g is a surjective polynomial function. But the tree $T_{ab|cd}$ is generically identifiable by Proposition 11, and thus invariable site parameters in $\mathbf{s}_{abcd}$ are generically identifiable by Proposition 12. As these coincide with the invariable site parameters in $\mathbf{s}$, and generic conditions on $\mathbf{s}_{abcd}$ imply generic conditions on $\mathbf{s}$, the invariable site parameters are generically identifiable for the full n-taxon model.

As an n-taxon binary tree topology is determined by the collection of all induced quartet tree topologies, one can now see that T is generically identifiable. Alternately, using the identified invariable site parameters, and assuming the additional generic condition that $\delta \neq 1$, note that

$$P_{\mathrm{GM}} = \frac{1}{(1-\delta)}(P - \delta P_{\mathrm{I}})$$

is a joint distribution arising from general Markov parameters. Thus generic identifiability of the tree can also by obtained from Steel's result for the GM model [2] applied to P_{GM}.

The generic identifiability of the remaining numerical parameters follows from Chang's argument [3] applied to P_{GM}. Chang's approach also indicates the cardinality of the generic fiber is $\kappa!(|V| - n)$ due to the label swapping phenomenon. □

Author's personal copy

5. Estimating invariable sites parameters

The concrete result in Proposition 12 gives explicit rational formulas for recovering parameters relating to invariable sites from the joint distribution. These can be viewed as generalizations of the formulas found in [11] for group-based models. As [11] develops the group-based model formulas into a heuristic means of estimating the invariable site parameters from data without performing a full Maximum Likelihood fit of data to a tree under a $\mathcal{M}+I$ model, one might suspect the formulas of Proposition 12 could be used similarly without the need to assume $\mathcal{M}$ was group-based, or approximately group-based. We emphasize that however useful such an estimate might be, it would not be intended to replace a more statistical but time-consuming computation, such as obtaining the Maximum Likelihood estimates for these parameters. (See [13].)

However, it is by no means obvious how to use these formulas well even for a heuristic estimate. First, for a 4-taxon tree we have many choices for the matrix B, in fact

$$\binom{\kappa^2-\kappa}{\kappa}^2$$

of them, so even for $\kappa=4$, there are 245 025 basic sets of the formulae. Moreover, while these simple formulae emerged from our method of proof, one could in fact modify them by adding to any of them a rational function whose numerator is a phylogenetic invariant for the GM + I model, and whose denominator is not. Since the invariant vanishes on any joint distribution arising from the model, the resulting formulae will still recover invariable site information for generic parameters. Thus there are actually infinitely many formulas for recovering invariable site parameters.

One can nonetheless consider simple averaging schemes using only the basic formulas of Proposition 12 and find that on simulated data they perform quite well at approximately recovering invariable site parameters from empirical distributions. However, averaging the large number of formulas give here, and then also averaging over a large sample of quartets, as is proposed in [11], is more time consuming than one might wish for a fast heuristic. Moreover, one must be aware that the denominator in these formulas may vanish on an empirical distribution – it is certain to be non-zero only for true distributions for GM + I arising from generic parameters.

Nonetheless, it would be of interest to develop versions of these formulas with good statistical estimation properties, as the GM + I model encompasses models such as the GTR + I model which is often preferred in biological data analysis to group-based + I models.

Appendix A. Code for computational algebra software

The following code is also available on the authors' websites.

A.1. Computation for Proposition 6

To show the variety has dimension 13, we execute the following Maple code:

Author's personal copy


```
pa:=Matrix([[p,1−p]]); Mae:=Matrix([[1−a,a],[r,1−r]]);
Meb:=Matrix([[1−b,b],[s,1−s]]); Mef:= Matrix([[1−e,e],[t,1−t]]);
Mfc:=Matrix([[1−c,c],[u,1−u]]); Mfd:= Matrix([[1−d,d],[v,1−v]]);
P:=Array(1..2,1..2,1..2,1..2);
for i from 1 to 2 do for j from 1 to 2 do for k from 1 to 2 do for l from 1 to 2 do
  P[i,j,k,l]:=0;
  for m from 1 to 2 do for n from 1 to 2 do
    P[i,j,k,l]:=P[i,j,k,l]+pa[l,i]*Mae[i,m]*Meb[m,j]*Mef[m,n]*Mfc[n,k]*
      Mfd[n,l];
  od;od;
  P[i,j,k,l]:=(1−w)*P[i,j,k,l];
od;od;od;od;
P[1,1,1,1]:=P[1,1,1,1]+w*q: P[2,2,2,2]:=P[2,2,2,2]+w*(1−q):
Q:=ListTools[Flatten](convert(P,listlist)):
J:=VectorCalculus[Jacobian](Q,[a,b,c,d,e,r,s,t,u,v,p,q,w]):
K:=subs(a=1/3,b=1/5,c=1/7,d=1/11,e=1/13,r=1/17, s=1/19,t=1/23,u=1/29,
   v=1/31,p=1/3,q=1/5,w=1/7,J):
LinearAlgebra[Rank](K);
```

Using Singular [21], we complete the proof:

```
LIB "matrix.lib"; LIB "primdec.lib";
ring r=0, (p0,p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,p11,p12,p13,p14,p15),dp;
// Define matrix flattening F_{ab|cd} and polys fs, f1, f2
matrix Fab[4][4]=p0,p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,p11,p12,p13,p14,p15;
matrix UR[3][3]=submat(Fab,1..3,2..4); matrix LL[3][3]=submat(Fab,2..4,
  1..3);
poly f1=det(UR); poly f2=det(LL);
poly fs=p0+p1+p2+p3+p4+p5+p6+p7+p8+p9+p10+p11+p12+p13+p14+p15−1;
ideal I=fs,f1,f2; // define ideal I
dim(std(I));  // compute dimension of r/I
primdecGTZ(I);  // compute primary decomposition of I to show prime
```

A.2. Computation for intersections of $V_{ab|cd}, V_{ac|bd}, V_{ad|bc}$

Continuing the Singular session above, we execute the following:

```
/* Define ideals Iac, Iad corresponding to two alternative tree
   topologies for 4-taxon trees. (So, I = Iab in this notation.)  */
// Flattening for ac|bd split
matrix Fac[4][4]=p0,p1,p4,p5,p2,p3,p6,p7,p8,p9,p12,p13,p10,p11,p14,p15;
poly f3=det(submat(Fac,1..3,2..4)); poly f4=det(submat(Fac,2..4,1..3));
ideal Iac=fs,f3,f4;
```

Author's personal copy


```
// Flattening for ad|bc split
matrix Fad[4][4]=p0,p2,p4,p6,p1,p3,p5,p7,p8,p10,p12,p14,p9,p11,p13,p15;
poly f5=det(submat(Fad,1..3,2..4)); poly f6=det(submat(Fad,2..4,1..3));
ideal Iad=fs,f5,f6;
reduce(f1,std(Iac)); // non-zero answer shows f1 not in Iac
reduce(Iac,std(I)); // non-zero shows f3,f4 not in I
ideal J=I,Iac; dim(std(J)); // show dim is 11
ideal K=J,Iad; dim(std(K)); // show dim is 11
primdecGTZ(K);   // show K prime, and thus ideal for star tree
```

A.3. Computation of 2-state GM + I ideal, 4-taxon trees, using Proposition 9

The following Singular code performs the needed elimination for a binary tree:

```
ideal Igm=minor(Fab,3);
// Eliminate the 'diagonal' variables
ideal Igmi=eliml(Igm,p0*p15);
```

For the star tree, the 2-state GM ideal is known from [16]. Thus elimination can be used to find GM + I invariants. We also show this result agrees with `K` above.

```
ideal Igm=minor(Fab,3),minor(Fac,3),minor(Fad,3);
// Eliminate the 'diagonal' variables
ideal Igmi=eliml(Igm,p0*p15),fs;
reduce(K,std(Igmi)); // all 0's indicates ideal containment
reduce(Igmi,std(K)); // all 0's indicates ideal containment
```

References

[1] P. Buneman, The recovery of trees from measures of dissimilarity, in: Mathematics in the Archeological and Historical Sciences, Edinburgh University Press, Edinburgh, 1971, p. 387.

[2] M. Steel, Recovering a tree from the leaf colourations it generates under a Markov model, Appl. Math. Lett. 7 (2) (1994) 19.

[3] J.T. Chang, Full reconstruction of Markov models on evolutionary trees: identifiability and consistency, Math. Biosci. 137 (1) (1996) 51.

[4] X. Gu, W.-H. Li, A general additive distance with time-reversibility and rate variation among sites, PNAS 93 (1996) 4671.

[5] P.J. Waddell, M. Steel, General time-reversible distances with unequal rates across sites: mixing Γ and inverse Gaussian distributions with invariant sites, Mol. Phylo. Evol. 8 (3) (1997) 398.

[6] J.S. Rogers, Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution, Syst. Biol. 50 (5) (2001) 713.

[7] C. Ané, personal communication (2005).

[8] E.S. Allman, J.A. Rhodes, The identifiability of tree topology for phylogenetic models, including covarion and mixture models, J. Comput. Biol. 13 (5) (2006) 1101. <arXiv:q-bio.PE/0511009>.

[9] C. Tuffley, M. Steel, Modeling the covarion hypothesis of nucleotide substitution, Math. Biosci. 147 (1) (1998) 63.

Author's personal copy

[10] E. Baake, What can and what cannot be inferred from pairwise sequence comparisons? Math. Biosci. 154 (1) (1998) 1.
[11] M. Steel, D. Huson, P.J. Lockhart, Invariable sites models and their uses in phylogeny reconstruction, Syst. Biol. 49 (2) (2000) 225.
[12] D.H. Huson, D. Bryant, Application of phylogenetic networks in evolutionary studies, Mol. Biol. Evol. 23 (2) (2006) 254. <http://www.splitstree.org>.
[13] V. Jayaswal, J. Robinson, L. Jermiin, Estimation of phylogeny and invariant sites under the general Markov model of nucleotide sequence evolution, Syst. Biol. 56 (2) (2007) 155.
[14] M. Steel, L. Székely, M. Hendy, Reconstructing trees from sequences whose sites evolve at variable rates, J. Comput. Biol. 1 (2) (1994) 153.
[15] E.S. Allman, J.A. Rhodes, Phylogenetic invariants for the general Markov model of sequence mutation, Math. Biosci. 186 (2003) 113.
[16] E.S. Allman, J.A. Rhodes, Phylogenetic ideals and varieties for the general Markov model, Adv. Appl. Math. To appear, <arXiv:math.AG/0410604>.
[17] J.A. Cavender, J. Felsenstein, Invariants of phylogenies in a simple case with discrete states, J. Class. 4 (1987) 57.
[18] J. Lake, A rate independent technique for analysis of nucleic acid sequences: evolutionary parsimony, Mol. Bio. Evol. 4 (2) (1987) 167.
[19] E.S. Allman, J.A. Rhodes, Phylogenetic invariants, in: O. Gascuel, M. Steel (Eds.), New Mathematical Models of Evolution, Oxford University, 2007, pp. 108–147.
[20] T.R. Hagedorn, Determining the number and structure of phylogenetic invariants, Adv. Appl. Math. 24 (1) (2000) 1.
[21] G.-M. Greuel, G. Pfister, H. Schönemann, SINGULAR 3.0, A Computer Algebra System for Polynomial Computations, Centre for Computer Algebra, University of Kaiserslautern, <http://www.singular.uni-kl.de>, 2005.